Bayes linear analysis for graphical models: The geometric approach to local computation and interpretive graphics

M. Goldstein
University of Durham

D.J. Wilkinson[*]
University of Newcastle

December 12, 1997

Abstract

This paper concerns the geometric treatment of graphical models using Bayes linear methods. We introduce Bayes linear separation as a second order generalised conditional independence relation, and Bayes linear graphical models are constructed using this property. A system of interpretive and diagnostic shadings are given, which summarise the analysis over the associated moral graph. Principles of local computation are outlined for the graphical models, and an algorithm for implementing such computation over the junction tree is described. The approach is illustrated with two examples. The first concerns sales forecasting using a multivariate dynamic linear model. The second concerns inference for the error variance matrices of the model for sales, and illustrates the generality of our geometric approach by treating the matrices directly as random objects.

1 Introduction

The Bayes linear approach is motivated by both practical and theoretical considerations. The practical considerations arise from the need to develop methods of Bayesian analysis based on partial prior specifications for problems which are sufficiently complex that we are unable either to make a full prior specification or to carry out the full posterior analysis. The theoretical considerations follow from the value in creating an essentially geometric approach to belief specification and inference, which treats general random objects, for example random matrices, in a formally similar way to simple random quantities.

In this paper, we describe the treatment of Bayes linear graphical models. In particular, we consider how to adjust beliefs over such a model, and how to display the results of such adjustments. We therefore develop the general geometric approach to local computation, which allows us to analyse very large models. This approach is linked to a novel collection of interpretive and diagnostic graphical displays for the moral graph and the junction tree associated with the graphical model, which both display the flow of information across the model and also identify possible discrepancies between the model and observations. The formal approach that

[*] Address for correspondence: Dr. D.J. Wilkinson, Department of Statistics, University of Newcastle, Newcastle upon Tyne NE1 7RU, England. Email: d.j.wilkinson@newcastle.ac.uk WWW: http://www.ncl.ac.uk/~ndjw1/

we describe is quite general and may be used to analyse graphical models for general finite or infinite collections of random objects defined over any corresponding inner product space.

We proceed as follows. We begin by summarising the aspects of Bayes linear methodology that we shall require. We consider the interpretation of the methodology, including a discussion of the geometric structure underlying the approach. We then discuss Bayes linear separation which acts as a generalised conditional independence property, and therefore forms the basis for the construction of Bayes linear graphical models. We describe a system of interpretive and diagnostic shadings which represent the passage of information around the associated moral graph and junction tree. The geometric basis for local computation within the model is then discussed, and a suitable algorithm to implement such computations is described for the associated junction tree. The approach is illustrated with two examples. The first example concerns prediction for vectors of sales data generated from a multivariate dynamic linear model. The second example concerns inferences about the variance matrices underlying the dynamic linear model in the first example, and exploits the generality of the geometric formalism in order to simplify the analysis by treating the variance matrices directly as matrix objects within an appropriate inner product space.

2 Bayes linear graphical models

2.1 Bayes linear adjustment

In this section, we describe those properties of a Bayes linear analysis that we shall require: for an overview of Bayes linear analysis, see Goldstein (1998); for a discussion of Bayes linear computing, see Goldstein & Wooff (1995). We have a collection, C, of random quantities. For each quantity, we specify a prior mean and variance, and between each pair of quantities we specify a prior covariance. These specifications are made directly, treating expectation, rather than probability, as the primitive quantity; see, for example, the development in de Finetti (1974). The *adjusted expectation* of a random quantity X, given a collection $D = \{D_1, \ldots, D_k\}$, written $\mathrm{E}_D(X)$, is the linear combination $\mathrm{E}_D(X) = \sum_{i=0}^k h_i D_i$ which minimises $\mathrm{E}((X - \sum_{i=0}^k h_i D_i)^2)$, over all collections $h = (h_0, h_1, \ldots, h_k)$, where D_0 is the unit constant, i.e. $D_0 = 1$.

If B, D are finite vectors of random quantities, then the matrix representation for $\mathrm{E}_D(B)$, the vector of adjusted expectations for the elements of B by D, is

$$\mathrm{E}_D(B) = \mathrm{E}(B) + \mathrm{Cov}(B, D)(\mathrm{Var}(D))^\dagger(D - \mathrm{E}(D)), \tag{1}$$

where $(\mathrm{Var}(D))^\dagger$ is a generalised inverse of $\mathrm{Var}(D)$. We define the adjusted version of the vector B given D, denoted $[B/D]$, to be the residual vector $[B/D] = B - \mathrm{E}_D(B)$. Properties of the adjusted vector are

$$\mathrm{E}([B/D]) = 0,\ \mathrm{Cov}(\mathrm{E}_D(B), [B/D]) = 0, \tag{2}$$

In particular, as $B = \mathrm{E}_D(B) + [B/D]$, we have $\mathrm{Var}(B) = \mathrm{Var}(\mathrm{E}_D(B)) + \mathrm{Var}([B/D])$. We call $\mathrm{RVar}_D(B) = \mathrm{Var}(\mathrm{E}_D(B))$ the *resolved variance matrix* for B by D, and $\mathrm{Var}_D(B) = \mathrm{Var}([B/D])$ the *adjusted variance matrix*, for B by D. The matrix representation is

$$\mathrm{RVar}_D(B) = \mathrm{Cov}(B, D)(\mathrm{Var}(D))^\dagger\mathrm{Cov}(D, B),\ \mathrm{Var}_D(B) = \mathrm{Var}(B) - \mathrm{RVar}_D(B) \tag{3}$$

$$\mathrm{Cov}_D(B,C) = \mathrm{Cov}([B/D],[C/D]) = \mathrm{Cov}(B,C) - \mathrm{Cov}(B,D)(\mathrm{Var}(D))^{\dagger}\mathrm{Cov}(D,C), \tag{4}$$

When we evaluate the adjusted mean and variance matrix for a collection of random quantities B, we also evaluate the corresponding adjustments for each finite linear combination of the elements of B. We denote by $\langle B\rangle$ the linear space of all such finite linear combinations. A useful summary of the adjustment over $\langle B\rangle$ is the *resolution transform* over $\langle B\rangle$ induced by D, defined, for each $Y \in \langle B\rangle$, as $T_{D(B)}(Y) = \mathrm{E}_B(\mathrm{E}_D(Y))$; see Goldstein (1981). If the dimension of $\mathrm{Var}(B)$ is $k_B < \infty$, then $T_{D(B)}$ has k_B mutually uncorrelated eigenvectors, $Z_1, ..., Z_{k_B}$, each normed to prior variance 1, with corresponding eigenvalues $1 \geq \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{k_B} \geq 0$. The non-zero eigenvalues of $T_{D(B)}$ and $T_{B(D)}$ are the same. For any $Y \in \langle B\rangle$, we have $\mathrm{RVar}_D(Y) = \sum_i \lambda_i (\mathrm{Cov}(Y, Z_i))^2$. Therefore, the eigenvectors of $T_{D(B)}$ identify the directions in $\langle B\rangle$ over which we expect a substantial reduction in uncertainty by linear fitting on D. In particular, the largest and smallest eigenvalues identify the maximum and minimum reductions in variance over the elements of $\langle B\rangle$. If we choose k_B mutually uncorrelated elements of $\langle B\rangle$, each with prior variance 1, then $\mathrm{trace}(T_{D(B)})$ is the sum of the resolved variances for each element. The matrix representation for $T_{D(B)}$ is

$$T_{D(B)} = (\mathrm{Var}(B))^{\dagger}\mathrm{RVar}_D(B). \tag{5}$$

Now suppose that we first observe collection D_1 and then observe $D_2 = \{D_{21}, ..., D_{2s}\}$. We define the partial expectation for X given D_2 adjusted by D_1 as the linear combination $\mathrm{E}_{[D_2/D_1]}(X) = \sum_{i=1}^{s} h_i [D_{2i}/D_1]$ minimizing $\mathrm{E}((X - \sum_{i=1}^{s} h_i [D_{2i}/D_1])^2)$ over all choices of h. The vector $\mathrm{E}_{[D_2/D_1]}(B)$ may be evaluated as

$$\mathrm{E}_{[D_2/D_1]}(B) = \mathrm{Cov}(B,[D_2/D_1])\mathrm{Var}([D_2/D_1])^{\dagger}[D_2/D_1] = \mathrm{Cov}_{D_1}(B,D_2)\mathrm{Var}_{D_1}(D_2)^{\dagger}[D_2/D_1] \tag{6}$$

The partial resolved variance matrix for B given D_2 adjusted by D_1 is

$$\mathrm{RVar}_{[D_2/D_1]}(B) = \mathrm{Var}(\mathrm{E}_{[D_2/D_1]}(B)) = \mathrm{Cov}_{D_1}(B,D_2)\mathrm{Var}_{D_1}(D_2)^{\dagger}\mathrm{Cov}_{D_1}(D_2,B) \tag{7}$$

Denoting the corresponding adjusted expectation vector and resolved variance matrix for B given both D_1 and D_2 as $\mathrm{E}_{D_1+D_2}(B)$, $\mathrm{RVar}_{D_1+D_2}(B)$, we have

$$\mathrm{E}_{D_1+D_2}(B) = \mathrm{E}_{D_1}(B) + \mathrm{E}_{[D_2/D_1]}(B) \tag{8}$$

$$\mathrm{RVar}_{D_1+D_2}(B) = \mathrm{RVar}_{D_1}(B) + \mathrm{RVar}_{[D_2/D_1]}(B) \tag{9}$$

$$T_{(D_1+D_2)(B)} = T_{D_1(B)} + T_{[D_2/D_1](B)} \tag{10}$$

$$[B/(D_1+D_2)] = [[B/D_1]/[D_2/D_1]] \tag{11}$$

2.2 Interpretations

From a Bayesian viewpoint, adjusted expectation is a simple and tractable approximation to a full Bayes posterior expectation, based on limited prior specification, which is exact in certain important special cases. For example, if the D is a collection of indicator functions over a partition of events, then adjusted expectation given D is numerically identical to conditional expectation given the value of D. Bayes linear analysis is also exact for Gaussian models, and one feature of interest in the Bayes linear approach is to observe how much of the simplicity of the Gaussian analysis may be preserved without making the Gaussian assumptions.

Further, if we view expectation as primitive, then adjusted expectation generalises the notion of conditional expectation from indicator functions to general random quantities. From a fully subjectivist viewpoint, adjusted expectation is a prior inference, based on the limited prior specification, for the posterior expectation having seen the data; see Goldstein (1997).

More generally, the Bayes linear approach is relevant whenever the natural formalism of interest is not the Boolean space of outcomes of events but the linear space corresponding to combinations of the random quantities. This formalism has various technical advantages; for example, we may handle finite and infinite collections of quantities in identical fashion, and we may therefore represent by a single node on a graphical model objects such as an infinite exchangeable collection of potential data observations or the infinite collection of all random variables defined as measurable functions on a particular continuous sample space. More importantly, this formalism allows us to extend our analysis to quite general objects; for example, we shall illustrate how we might pass partially specified beliefs about a collection of variance matrices around a graphical structure, using the same formalism that we shall apply to simple random quantities.

The formal structure is as follows. We have a (finite or infinite) collection, C, of quantities of interest; for example, C might consist of random quantities, or it might consist of random $r \times r$ matrices. We form the linear space $\langle C \rangle$ of finite linear combinations of the elements of C. We then define an inner product $(\cdot, \cdot)$ over C to express some aspect of our prior beliefs over C. If C consisted of random quantities, then we might choose $(X, Y) = \mathrm{E}(XY)$, while if C was a collection of random matrices then we might choose $(X, Y) = \mathrm{E}(\mathrm{trace}(X^T Y))$; see Wilkinson & Goldstein (1996). We denote the closure of any linear subspace, $\langle B \rangle$ say, as $[B]$. Within this formalism, $\mathrm{E}_D(\cdot)$ is the projection operator from $[C]$ into $[D]$, and $\mathrm{Var}_D(\cdot)$ is the squared orthogonal distance to $[D]$. Further, $S_D = I - T_D$, where I is the identity transform and T_D is the resolution transform for D, is the unique self adjoint transform over $[C]$ which transforms the inner product $(\cdot, \cdot)$ into the adjusted inner product $(X, Y)_D = ([X/D], [Y/D])$, as for any $X, Y \in [C]$, we have $(X, S_D(Y)) = (X, Y)_D$.

For reasons of familiarity, we shall describe the approach to local computation using means, variances and covariances. However, our results extend immediately to general inner products over general linear spaces. At a formal level, our development therefore offers a general geometric approach to local computation.

2.3 Bayes linear separation

If A, B, C are three random vectors, then we say that B*separates* A and C, written $(A \perp\!\!\!\perp C)/B$, if

$$\mathrm{Cov}_B(A, C) = 0. \tag{12}$$

Within the more general geometric formalism, the condition $(A \perp\!\!\!\perp C)/B$ corresponds to the requirement on the adjusted inner product that $(X, Y)_B = 0, \forall X \in [A], Y \in [C]$.

Belief separation may be viewed as a *generalised conditional independence property* (Dawid 1979), namely a tertiary property on collections of objects which obeys the following three properties, for any collections B, C, D, E.

(C1) $(B \perp\!\!\!\perp C)/(C + D)$

(C2) $(B \perp\!\!\!\perp C)/D \Leftrightarrow (C \perp\!\!\!\perp B)/D$

(C3) $(B \perp\!\!\!\perp (C + D))/E$ implies and is implied by the pair of conditions (i) $(B \perp\!\!\!\perp D)/E$ and (ii) $(B \perp\!\!\!\perp C)/(D + E)$.

In Goldstein (1990), it is shown that belief separation is a generalised conditional independence property. Any tertiary relation obeying these properties will behave computationally as a conditional independence property; see Smith (1990). Therefore, graphical models based on belief separation will have many of the same qualitative properties as do probabilistic graphical models; see for example Pearl (1988).

Suppose that we have three belief structures A, B, C for which $(A \perp\!\!\!\perp C)/B$. Therefore, the covariance structure between the collections A and C is determined by the pair of covariance structures $\mathrm{Cov}(A, B)$ and $\mathrm{Cov}(B, C)$, as $(A \perp\!\!\!\perp C)/B$ implies that $(A - \mathrm{E}_B(A))$ is uncorrelated with C, i.e. that $\mathrm{Cov}(A_i, C_j) = \mathrm{Cov}(\mathrm{E}_B(A_i), C_j) \forall A_i \in A, C_j \in C$. In particular, if A, B, C are finite vectors with $(A \perp\!\!\!\perp C)/B$, then we have

$$\mathrm{Cov}(A, C) = \mathrm{Cov}(A, B)(\mathrm{Var}(B))^{\dagger}\mathrm{Cov}(B, C) \tag{13}$$

In our development, we shall adjust collections of quantities progressively, introducing pieces of information in a stepwise manner. We identify the circumstances where belief separation is preserved under adjustment as follows.

Theorem 1 *Suppose that A, B, C are three belief structures for which $(A \perp\!\!\!\perp C)/B$. For any further belief structure D we have that $([A/D] \perp\!\!\!\perp [C/D])/[B/D]$ if and only if $(A \perp\!\!\!\perp C)/(B + D)$. In particular, a sufficient condition for $([A/D] \perp\!\!\!\perp [C/D])/[B/D]$ is that $(A \perp\!\!\!\perp (C + D))/B$*

Proof The condition $([A/D] \perp\!\!\!\perp [C/D])/[B/D]$ is equivalent to the condition

$$(A - \mathrm{E}_D(A)) - \mathrm{E}_{[B/D]}((A - \mathrm{E}_D(A))) \perp (C - \mathrm{E}_D(C)) - \mathrm{E}_{[B/D]}((C - \mathrm{E}_D(C)))$$

which reduces to the condition $(A \perp\!\!\!\perp C)/(B + D)$. A sufficient condition to ensure that $(A \perp\!\!\!\perp C)/(B + D)$ is that $(A \perp\!\!\!\perp (C + D))/B$, from (C3). □

2.4 Bayes linear graphical models

We represent a collection of belief separations with a *Bayes linear graphical model*, based on a directed graph, with nodes $B_1, \ldots B_r$, say, where each node B_i represents a collection of random quantities. Certain nodes are joined by directed arrows. If a directed arc goes from node A to node B, then A is termed a parent of B, B is termed a child of A and A, B are said to be adjacent or neighbour nodes. We denote by $P(B)$ the set of parents of B. We say that a directed acyclic graph is a *directed (second-order) graphical model* if for any nodes B_i and B_j we have $(B_i \perp\!\!\!\perp B_j)/(P(B_i) + P(B_j))$. In order to specify second-order beliefs between all pairs of nodes, it is sufficient to specify the variance matrix for each node and covariance matrices between each parent-child pair of nodes, from (13).

An alternative way to represent a collection of belief separations is through an undirected graph, where each node represents a collection of random quantities, and certain pairs of nodes are joined by undirected arcs. We say that a collection of nodes B *separates* the collections A and C of nodes on the graph, if every path from a node in A to a node in C passes through a node in B. We say that an undirected graphical model has the *second-order global Markov property* if, whenever B separates A and C on the graph, then $(A \perp\!\!\!\perp C)/B$. In particular, we may construct the *moral graph* for a directed graphical model by drawing an arc between any two nodes which are parents of the same child node and which are not currently joined by an arc, and dropping all arrows. The moral graph is second-order global Markov, as follows; see for example Lauritzen, Dawid, Larsen & Leimer (1990), section (6).

Theorem 2 *For any three collections of nodes A, B, C, within a directed graphical model, construct the moral graph on A, B, C and all ancestors. If B separates A from C on this graph, then $(A \perp\!\!\!\perp C)/B$.*

As we receive data, certain of the nodes become known. The following result allows us to remove observed nodes from the moral graph.

Theorem 3 *Suppose that an undirected graphical model is second-order global Markov. Choose any node, D say, and remove node D and all arcs entering D from the diagram. The resulting diagram is second-order global Markov for the collection of beliefs resulting from adjusting all quantities by D.*

Proof Remove node D and all arcs entering D from the graph. Suppose that on the new graph all paths from A to B pass through C. Therefore, all paths from A to B on the original diagram pass through C or D, so that $(A \perp\!\!\!\perp B)/(C + D)$. This implies, from theorem 1, that $([A/D] \perp\!\!\!\perp [B/D])/[C/D]$, so that separation on the modified graph corresponds to the separation of adjusted beliefs as required. □

2.5 Interpretive and diagnostic shadings on the moral graph

In Goldstein, Farrow & Spiropoulos (1993), a system of interpretive and diagnostic shadings was described for the analysis of directed second order graphical models. For representing and analysing the progressive effects of local computations, we suggest an alternative display on the moral graph, which serves as a graphical analogue to the formal calculations. We have observed, in theorem 3, that the effect of fully observing a node is to remove that node, and the corresponding arcs, from the graph. Therefore, a natural graphical display is to shrink each node as partial information is received, in proportion to the reduction in variance, which we do as follows.

Each node B has standard initial width, which may be taken to be proportional to the dimension, k_B say, of $\mathrm{Var}(B)$. If we adjust the model by D_1, then we remove an outer ring from each node B with area proportional to the trace of the resolution transform $T_{D_1(B)}$, which represents the overall reduction in variance over $\langle B \rangle$.

Having adjusted by D_1, suppose that we further adjust all nodes by D_2, then D_3, and so forth. Let $D(s) = D_1 + \cdots + D_s$. As we adjust by each D_r, we remove a further outer ring from each node. The area of the ring is proportional to the additional resolved variance from the latest adjustment, namely the trace of $T_{[D_r/D(r-1)](B)}$, so that, from (10), the total area removed after adjustment by $D_1, ..., D_r$ is proportional to trace $T_{D(r)(B)}$. We leave a faint outer diameter to show the original size of each node, so that, when a node has been fully observed, it leaves a ghost image behind.

As these displays offer simple visual summaries of the effects of belief adjustments, they are of particular use at the design stage when we are choosing which quantities to observe. While automatic searches for good designs may be useful, any design for a complex high-dimensional system will have many implications which are not captured by any tractable search criterion. Therefore, a sensible way to select good designs is to use a broad range of criteria as screening devices to rule out clearly inadequate designs, and then to use careful graphical comparisons to select between the leading choices for the design. The qualitative understanding that we gain by this approach should allow us to select a design whose strengths and weaknesses are clearly understood. Further, this graphical representation forms the basis for diagnostic evaluation of the model, as follows.

As we observe data $D_1 = d_1$, we evaluate the adjusted expectation, $\mathrm{E}_{d_1}(Y)$, for each Y. There are various diagnostic measures to assess our observed changes in expectation. A simple measure is $H(d_1) = \max_{Y \in \langle B \rangle} (\mathrm{E}_{d_1}(Y) - \mathrm{E}(Y))^2 / \mathrm{Var}(Y)$. The prior expected value of $H(D_1)$ is equal to the trace of $T_{D_1(B)}$, so the comparison of the two quantities is a simple way to assess the consistency of our prior specification with our observations; for discussion of the interpretation and evaluation of this measure, see Goldstein (1988). Therefore, a natural way to show the diagnostic information on the graph is to display the ratio $R(d_1) = H(d_1)/\mathrm{trace}(T_{D_1(B)})$, by shading a proportion of the area which has been removed.

A simple shading is to shade nothing if $R(d_1) = 1$, to use red, or dark, shading to show large values of $R(d_1)$ and blue, or light shading for values of $R(d_1)$ less than one. The amount that we shade depends on whether we want only to highlight extreme values, for example if we are trying to construct a simplified, approximate diagram for a complex system, or if we want to draw attention even to minor discrepancies, for example if we have been successfully using versions of this diagram to forecast over a period of time. A simple general purpose choice of shading follows from comparing the value of $R(d_1)$ to the tail area corresponding to some choice of simple approximating distribution. For example, if all the elements of D_1 are normally distributed, then $\mathrm{Var}(R(D_1)) = V_\lambda = 2 \sum_{i=1}^{r(B)} \lambda_i^2 / (\sum_{i=1}^{r(B)} \lambda_i)^2$. As a simple approximation to the distribution of $R(D_1)$ we might choose a gamma distribution with mean 1 and variance V_λ. Note that $V_\lambda = 2\mathrm{trace}(T_{D(B)}^2)/(\mathrm{trace}(T_{D(B)}))^2$, and so may be evaluated without finding the full eigen decomposition of $T_{D(B)}$.

If we further observe data $D_2 = d_2$, then $H([d_2/d_1]) = \max_{Y \in \langle B \rangle} (\mathrm{E}_{d_1+d_2}(Y) - \mathrm{E}_{d_1}(Y))^2 / \mathrm{Var}(Y)$ has expectation equal to the trace of $T_{[D_2/D_1](B)}$, so we may shade the second segment of the node to display the diagnostic ratio $R([d_2/d_1]) = H([d_2/d_1])/\mathrm{trace}(T_{[D_2/D_1](B)})$. Having successively observed $D_i = d_i, i = 1, ..., r$, or equivalently observing $D(r) = d(r)$ we may either choose to display (i) the *partial adjustment graph*, in which we display each ratio $R([d_i/d(i-1)])$ in the corresponding ring of the node, or (ii) the *total adjustment graph* in which we only display the ratio for the overall adjustment, $R(d(r))$.

The inner ring of node B always corresponds to the final portion of variation which would be resolved if all the quantities in B were observed. If we observe B, having first observed $d_1, ..., d_r$, then we may choose to assess and display the quantity $R([b/d(r)])$ in the central region.

In the special case where the graph is a tree, namely when there is a unique path between any two nodes, then we can follow the passage of each piece of information around the arcs of the tree, which we do as follows. If arc A_{UV} joins nodes U, V, then the arc has an initial thickness proportional to trace$(T_{U(V)})$, or equivalently, trace$(T_{V(U)})$. This is the maximal amount of information which can pass between the two nodes, after which the nodes will no longer be joined.

At stage r, suppose that D_r is nearer to U than V. Therefore, trace$(T_{[D_r/D(r-1)](U)}) \geq$ trace$(T_{[D_r/D(r-1)](V)})$. trace$(T_{[D_r/D(r-1)](V)})$is the information that has passed along the arc, while trace$(T_{[D_r/D(r-1)](U)})$ − trace$(T_{[D_r/D(r-1)](V)})$ is the information that is lost between the nodes. We reduce the arc thickness by an amount proportional to trace$(T_{[D_r/D(r-1)](V)})$. As with the nodes, we leave a faint outer edge for each arc so that, when two nodes become disconnected, a ghost arc remains. As we observe the data, we may make diagnostic shadings on the arcs in an analogous way to the node shadings, namely to shade in proportion to the value of $R([d_r/d(r-1)])$ for V. We may construct and display either or both of (i) the *stepwise adjustment graph*, in which we display each ratio $R([d_i/d(i-1)])$ in the corresponding section of the arc, and (ii) *the current adjustment graph* in which we renormalise after each message has passed, so that each non-zero arc always has a fixed thickness which, after $D(r-1)$ has been observed, is proportional to trace$(T_{[U/D(r-1)]([V/D(r-1)])})$, and we only shade the current adjustment, in proportion to trace$(T_{[D_r/D(r-1)]([V/D(r-1)])})$. This latter shading may be used to trace the propagation of individual errors and allows us to focus attention solely on the effect upon the current state of the system of adjustment by the latest collection of observations (whose location we would usually identify on the graph).

3 Local computation

We now describe how to carry out local computations over a Bayes linear graphical model. We first describe geometrically why such computation is always possible, and then give an algorithm based on the junction tree for local computation.

3.1 Elements of local computation

The essential requirement for second order local computation is as follows. If B separates A from C, then we want to adjust C by A purely in terms of computations on the pair A and B followed by computations on the pair B and C. We now show how each ingredient of the adjustment may be assessed in this stepwise fashion.

Theorem 4 *If* $(A \perp\!\!\!\perp C)/B$, *then*

(i)

$$\mathrm{E}_A(C) = \mathrm{E}_A(\mathrm{E}_B(C)) \tag{14}$$

(ii)

$$\mathrm{Var}_A(C) = \mathrm{Var}_B(C) + \mathrm{Var}_A(\mathrm{E}_B(C)) \tag{15}$$

(iii)

$$\mathrm{Cov}_A(B, C) = \mathrm{Cov}_A(B, \mathrm{E}_B(C)) \tag{16}$$

(iv)

$$T_{A(C)}(Y) = \mathrm{E}_C(T_{A(B)}(\mathrm{E}_B(Y))), \forall Y \in \langle C \rangle. \tag{17}$$

Proof (i) follows as $\mathrm{E}_A(C) = \mathrm{E}_A(\mathrm{E}_{A+B}(C))$ and $\mathrm{E}_{A+B}(C) = \mathrm{E}_B(C)$. (ii) follows as $(C - \mathrm{E}_B(C))$ is uncorrelated with $(\mathrm{E}_B(C) - \mathrm{E}_A(C))$. (iii) follows as $(C - \mathrm{E}_B(C))$ is uncorrelated with $A + B$. (iv) follows as $T_{A(C)} = \mathrm{E}_C(\mathrm{E}_A(C)) = \mathrm{E}_C(\mathrm{E}_B(\mathrm{E}_A(\mathrm{E}_B(C)))) = \mathrm{E}_C(T_{A(B)}\mathrm{E}_B(C))$. □

Theorem 4 shows that, in any finite or infinite system, we may exploit belief separation to implement local computation. All that we require is an efficient way of passing around the graph the information which is required in order to make these assessments. This algorithm will be based on the junction tree for the graph.

3.2 Junction Trees

The junction tree for the moral graph may be created as follows; see Jensen (1996), for a general discussion of the construction and use of the junction tree for probabilistic propagation, and see the appendices of Dawid & Lauritzen (1993) for analysis of the properties of junction tree type constructions under generalised conditional independence.

(i) Create the moral graph, by joining all parents and dropping arrows. (ii) Triangulate the graph, by adding sufficient edges to ensure that there are no cycles of length 4 or more without a chord. (iii) Arbitrarily label any node as node 1. (iv) Carry out a maximum cardinality search: at each stage k, label as node k any one of the nodes on the graph with the largest number of labelled neighbours. (If and only if the graph is triangulated, at each stage when we label node k, all labelled neighbours of this node will be neighbours of each other.) (v) Order the cliques (the maximal sets of nodes which are all joined to each other). For each clique, note the highest labelled node, and label the cliques in the order of these values. (vi) Create the junction tree as follows. The nodes of the tree are the cliques. Each clique is joined to at most one of the lower numbered cliques as follows. From the above construction, it turns out that the intersection of the nodes in a clique and the nodes in all lower numbered cliques will be contained in at least one of the lower numbered cliques. Place a link between the clique and any one of the lower numbered cliques which contain the intersection.

If the original directed graph represents a second-order graphical model, then the junction tree is an undirected graph with the second-order global Markov property. There is at most one path between any two nodes on the graph. Further, if a node of the original graph is contained in two nodes on the junction tree, then it is contained in all nodes on the unique path between these nodes. We also have the following property. Suppose that nodes A, B are adjacent on the junction tree. Let Z be the collection of nodes from the original graph which are in the intersection of A and B. Let U be the collection of nodes in A but not in Z, and let V be the nodes in B but not in Z. Then we must have $(U \perp\!\!\!\perp V)/Z$. In particular, the covariance between adjacent nodes on the junction tree may be derived from the covariances within each node, as from (13)

$$\mathrm{Cov}(U,V) = \mathrm{Cov}(U,Z)(\mathrm{Var}(Z))^{\dagger}\mathrm{Cov}(Z,V) \tag{18}$$

We now describe an algorithm for propagating belief adjustment around the junction tree which is based strictly on passing messages concerning the adjusted inner product between the data and all other quantities.

3.3 Local computation on the junction tree

Denote the nodes on the moral graph as $D_1, ..., D_m$. Denote the nodes of the junction tree as $J_1, ..., J_u$. We sequentially adjust beliefs over the whole graph as we observe collections $D_{(1)}, ..., D_{(k)}$, where each collection $D_{(i)}$ is contained in some node D_j on the moral graph, as follows. We firstly adjust each J_r by $D_{(1)}$. We next adjust each adjusted node $[J_r/D_{(1)}]$ by $[D_{(2)}/D_{(1)}]$. From (11), this is equivalent to the adjustment $[J_r/(D_{(1)} + D_{(2)})]$. We now adjust each $[J_r/(D_{(1)} + D_{(2)})]$ by $[D_{(3)}/(D_{(1)} + D_{(2)})]$, and so forth. A convenient algorithm to carry out this sequence of adjustments is as follows. Although we present the algorithm in terms of expectations and variances, it will work in the same way for a general inner product as described in section 2.2.

At stage $(i-1)$, for each node J_r, we have evaluated $\mathrm{E}_{[i-1]}(J_r)$, and $\mathrm{Var}_{[i-1]}(J_r)$, the adjusted expectation vector and variance matrix for J_r given $D[i-1] = D_{(1)} + ... + D_{(i-1)}$. For each j, we have evaluated $T_{[i-1](D_j)}$, the resolution transform for D_j given $D[i-1]$. We also retain the original variance matrix $\mathrm{Var}(J_r)$ for each collection. We have constructed a valid junction tree for the second order global Markov graph $M[i-1]$ in which each node D_j has variance matrix $\mathrm{Var}_{[i-1]}(D_j)$, and for which each node which has zero adjusted variance matrix (for example, each node for which all the elements have been observed), and all arcs into each such node, has been removed from the graph. We now describe how to adjust beliefs given $D_{(i)}$

1. We first pass the adjusted covariance, given $D[i-1]$, between $D_{(i)}$ and all the other nodes around the junction tree. $D_{(i)}$ is contained in each of some connected sequence of nodes on the junction tree, so that $\mathrm{Cov}_{[i-1]}(D_{(i)}, J)$ is already determined for these nodes. For each other node in turn we proceed as follows.

   Suppose that we have already assessed $\mathrm{Cov}_{[i-1]}(D_{(i)}, J_r)$ for node J_r, and we wish to pass the covariance to adjacent node s. If node J_s has subsets U_s, W_{rs} and node J_r has subsets U_r, W_{rs} where U_s, U_r, W_{rs} are disjoint, then $(U_s \perp\!\!\!\perp U_r)/W_{rs}$. Therefore, from (13), we find $\mathrm{Cov}_{[i-1]}(D_{(i)}, J_s)$ by

$$\mathrm{Cov}_{[i-1]}(D_{(i)}, U_s) = \mathrm{Cov}_{[i-1]}(D_{(i)}, W_{rs})\mathrm{Var}_{[i-1]}(W_{rs})^{\dagger}\mathrm{Cov}_{[i-1]}(W_{rs}, U_r).$$

2. Within each node J_s on the junction tree, we compute the partial mean and variance as

$$\mathrm{E}_{[i/(i-1)]}(J_s) = \mathrm{E}_{[D_{(i)}/D[i-1]]}(J_s) = \mathrm{Cov}_{[i-1]}(J_s, D_{(i)})\mathrm{Var}_{[i-1]}(D_{(i)})^{\dagger}(D_{(i)} - \mathrm{E}_{[i-1]}(D_{(i)})) \tag{19}$$

$$\mathrm{RVar}_{[i/(i-1)]}(J_s) = \mathrm{RVar}_{[D_{(i)}/D[i-1]]}(J_s) = \mathrm{Cov}_{[i-1]}(J_s, D_{(i)})\mathrm{Var}_{[i-1]}(D_{(i)})^\dagger\mathrm{Cov}_{[i-1]}(D_{(i)}, J_s) \qquad (20)$$

and for each node, D_j, on the moral graph, we evaluate the partial resolution transform matrix as

$$T_{[i/(i-1)](D_j)} = T_{[D_{(i)}/D[i-1]](D_j)} = \mathrm{Var}(D_j)^\dagger \mathrm{RVar}_{[i/(i-1)]}(D_j) \qquad (21)$$

As we need to use each prior inverse $\mathrm{Var}(D_j)^\dagger$, at each pass through the algorithm, it will often be efficient to store these inverses.

With the notation of step **(1)**, if we also pass the adjustments for W_{rs}, then we need only evaluate (19), (20), (21) for U_s, completing the resolved variance specification by

$$\mathrm{RVar}_{[i/(i-1)]}(U_s, W_{rs}) = \mathrm{Cov}_{[i-1]}(U_s, D_{(i)})\mathrm{Var}_{[i-1]}(D_{(i)})^\dagger\mathrm{Cov}_{[i-1]}(D_{(i)}, W_{rs})$$

3. We now update the adjusted mean and variance for each node J_s, from (8), (9), as

$$\mathrm{E}_{[i]}(J_s) = \mathrm{E}_{[i-1]}(J_s) + \mathrm{E}_{[i/(i-1)]}(J_s)$$

$$\mathrm{RVar}_{[i]}(J_s) = \mathrm{RVar}_{[i-1]}(J_s) + \mathrm{RVar}_{[i/(i-1)]}(J_s),\ \ \mathrm{Var}_{[i]}(J_s) = \mathrm{Var}(J_s) - \mathrm{RVar}_{[i]}(J_s)$$

and for each node, D_j, on the moral graph, we evaluate the resolution transform matrix, from (10) as

$$T_{[i](D_j)} = T_{[i-1](D_j)} + T_{[i/(i-1)](D_j)}$$

4. If we want to create the graphical representation for the adjustment, then we add the shading corresponding to the current stage of the adjustment to the moral graph and the junction tree. Using the shadings described in section 2.5, we maintain the record of the individual stages in the adjustment on the nodes of the moral graph, constructing either or both of the partial and total adjustment graphs, and we maintain a record of the current message passing on the arcs of the junction tree, constructing either or both of the stepwise and current adjustment graphs. As we often monitor the junction tree simply to detect problems in the computation, we will often be most concerned with the current adjustment graph. Thus, we are likely to inspect the current graphs for the first few such adjustments to check that the system appears to be working satisfactorily, and only to look at further problems when either certain diagnostic thresholds are exceeded for the arc diagnostics or we detect problems by inspecting the displays on the moral graph.

5. Each adjustment by $D_{(i)}$ preserves separations in the junction tree, by Theorem 1. Therefore, removing the collection $D_{(i)}$ from each node J_u on the junction tree where it appears, and removing any node on the junction tree which now has zero adjusted variance matrix, and all arcs into that node, the resulting structure is a junction tree when all beliefs are adjusted by $D[i]$. This junction tree corresponds to the second-order global Markov

undirected graph $M[i]$ obtained by removing all nodes, whose adjusted variance matrix is now zero, and all arcs into such nodes, from $M[i-1]$, by Theorem 3.

Note that this algorithm is well suited for parallel implementation. Firstly, the adjustment of each node by a given collection $D_{(i)}$ may proceed in parallel, as all that is required for the calculations within each node J is the adjusted covariance matrix, $\text{Cov}_{[i-1]}(D_{(i)}, J)$. Secondly, as the only message that we pass between neighbouring nodes is this adjusted covariance matrix, we may adjust the junction tree by several collections simultaneously. For example, if we want to adjust by both collections $D_{(u)}$ and $D_{(v)}$, then we may follow the above algorithm simultaneously to send both the adjustments around the tree. Some nodes will be first adjusted by $D_{(u)}$ and some by $D_{(v)}$, but the combined adjustment will be correct at each node. Belief updating in tree shaped networks via parallel message-passing between nodes is discussed in section 4.2 of Pearl (1988); see also section 7.2.2 of that work, where belief updating in Gaussian trees is also explained as a parallel message-passing algorithm.

4 Example: Local computation for a multivariate DLM

In Wilkinson & Goldstein (1997), a multivariate dynamic linear model (DLM) is considered, for the forecasting of sales of six different brands of soft drink. Modelling this, we have a sequence $X_1, X_2, \ldots$ of random vectors, each of length 6, such that $X_t = (X_{1t}, X_{2t}, \ldots, X_{6t})^T$. The component X_{it} represents the (unknown) sales of brand i at time t. The vectors of sales are modelled as a locally constant DLM as follows:

$$X_t = \Theta_t + \nu_t \quad \forall t \tag{22}$$

where

$$\Theta_t = \Theta_{t-1} + \omega_t \quad \forall t \tag{23}$$

where Θ_t is the 6 dimensional state vector for time t, and the qualitative form of the prior second-order specification is as follows:

$$\text{E}(\Theta_0) = e_0 \tag{24}$$

$$\text{E}(\nu_t) = \text{E}(\omega_t) = 0, \text{Var}(\Theta_0) = \Sigma, \tag{25}$$

$$\text{Var}(\nu_t) = V, \text{Var}(\omega_t) = W, \quad \forall t \tag{26}$$

$$\text{Cov}(\Theta_s, \nu_t) = \text{Cov}(\nu_s, \omega_t) = 0 \quad \forall s, t \tag{27}$$

$$\text{Cov}(\Theta_s, \omega_t) = 0 \quad \forall s < t \tag{28}$$

$$\text{Cov}(\omega_s, \omega_t) = \text{Cov}(\nu_s, \nu_t) = 0 \quad \forall s \neq t \tag{29}$$

Belief specifications are made as follows; see Wilkinson & Goldstein (1997) for more details.

$$e_0 \quad = \quad (119.05, 61.61, 3.00, 36.84, 35.41, 7.85)^T \tag{30}$$

$$\Sigma = \begin{pmatrix} 1161.61 & 213.51 & 12.57 & 77.60 & 24.27 & 26.89 \\ 213.51 & 147.75 & 3.93 & 34.29 & 13.12 & 4.64 \\ 12.57 & 3.93 & 18.78 & 2.70 & 1.44 & 3.67 \\ 77.60 & 34.29 & 2.70 & 72.48 & 12.20 & 6.32 \\ 24.27 & 13.12 & 1.44 & 12.20 & 28.98 & 3.50 \\ 26.89 & 4.64 & 3.67 & 6.32 & 3.50 & 24.83 \end{pmatrix} \tag{31}$$

$$V = \begin{pmatrix} 2420.36 & 387.33 & 20.39 & 165.27 & 44.56 & 58.61 \\ 387.33 & 263.85 & 3.85 & 71.51 & 23.48 & 3.27 \\ 20.39 & 3.85 & 30.79 & 3.58 & 1.26 & 5.99 \\ 165.27 & 71.51 & 3.58 & 139.72 & 23.12 & 11.33 \\ 44.56 & 23.48 & 1.26 & 23.12 & 50.01 & 4.78 \\ 58.61 & 3.27 & 5.99 & 11.33 & 4.78 & 44.21 \end{pmatrix} \tag{32}$$

$$W = \begin{pmatrix} 1112.49 & 272.47 & 22.52 & 66.45 & 31.56 & 27.84 \\ 272.47 & 195.50 & 11.53 & 30.07 & 18.51 & 15.37 \\ 22.52 & 11.53 & 29.64 & 5.54 & 4.67 & 6.28 \\ 66.45 & 30.07 & 5.54 & 78.91 & 14.04 & 8.03 \\ 31.56 & 18.51 & 4.67 & 14.04 & 40.50 & 7.32 \\ 27.84 & 15.37 & 6.28 & 8.03 & 7.32 & 32.97 \end{pmatrix} \tag{33}$$

In principle, one can now simply use the Kalman filter (or a variant thereof), in order to sequentially update the model with each observation in turn. However, in this case, there is particular interest in the residual structure of the model, and the effect of past, present and future observations on particular variables and residuals. This is useful for *inter alia*, diagnosing problems with the variance specification, spotting unusual patterns in the residual structure, and characterising surprising changes in the sales patterns. These questions may be tackled from a Kalman filtering perspective, provided that full forward and backward propagation of information is carried out after each time point observation. The local computation approach to be described illustrates this process from a Bayes linear viewpoint, using a framework which generalises to arbitrary model structures. It also illustrates the interpretive and diagnostic graphics which have been described, in the context of a DLM.

The moral graph for the problem is shown in Figure 1, together with some of the interpretive and diagnostic shadings already described. The shadings show the effect of the introduction of actual sales data for the first six time points, and will be described in greater detail shortly.

The key things to notice are that the moral graph nodes each represent a vector of six random quantities, that the moral graph is ready-triangulated and that the cliques of the graph all consist of three moral graph nodes. The cliques are arranged into two different types in the following manner.

$$A_i = \{\Theta_{i-1}, \Theta_i, \omega_i\} \tag{34}$$

$$B_i = \{X_i, \Theta_i, \nu_i\} \tag{35}$$

Figure 3 shows the junction (clique) tree for the problem, together with some interpretive and diagnostic shadings for the introduction of the sixth observation. In order to initialise the junction tree, we need to specify the expectation and covariance matrices for A_i and B_i. These are determined by the specifications already made, and take the following form.

$$\mathrm{E}(A_i) = \mathrm{E}(B_i) = (e_0^T, e_0^T, 0^T)^T \tag{36}$$

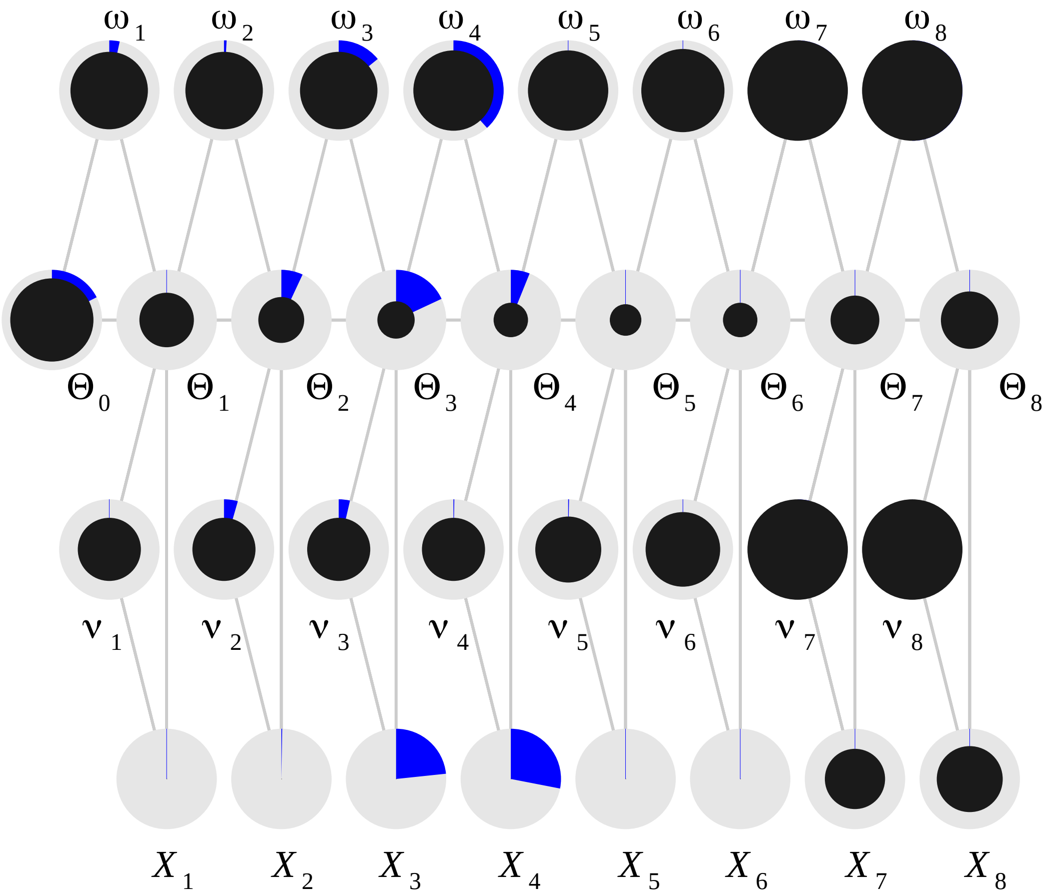

Figure 1: Global moral graph

$$\mathrm{Var}(A_i) = \begin{pmatrix} \Sigma + (i-1)W & \Sigma + (i-1)W & 0 \\ \Sigma + (i-1)W & \Sigma + iW & W \\ 0 & W & W \end{pmatrix} \tag{37}$$

$$\mathrm{Var}(B_i) = \begin{pmatrix} \Sigma + iW + V & \Sigma + iW & V \\ \Sigma + iW & \Sigma + iW & 0 \\ V & 0 & V \end{pmatrix} \tag{38}$$

Once the junction tree is initialised, local computation of adjustments may take place. Consider the introduction of data for X_1. The moral graph node, X_1, is contained in the junction tree node, B_1. $\mathrm{Cov}(X_1, B_1)$ is just the first six rows of $\mathrm{Var}(B_1)$, which is known, as the tree has been initialised. The covariance between X_1 and the other junction tree nodes can be found using the algorithm from section 3.3. B_1 has neighbours A_1 and A_2, and so $\mathrm{Cov}(X_1, A_1)$ and

$\text{Cov}(X_1, A_2)$ can be calculated using the fact that Θ_1 separates X_1 from both A_1 and A_2. Once these have been calculated, $\text{Cov}(X_1, B_2)$ can be calculated using $\text{Cov}(X_1, A_2)$ and the fact that Θ_2 separates X_1 from B_2. Propagation continues over the whole junction tree. Local adjustments may then be carried out at each node, before data on X_2 is introduced into B_2, *etc.*

Figure 1 shows the global picture after six observations. Consider (say) Θ_2; it is readily seen that over half of the area of the original node has been removed, indicating that over half of the original "variance" (uncertainty) associated with this node has been resolved. Note also that the circle representing remaining uncertainty has approximately half the diameter of the original circle, indicating that approximately half of the "standard deviation" associated with the node has been resolved. Some diagnostic shading is also present for this node. The approximate tail area scheme advocated in section 2.5 is used. Shadings are only made if the diagnostic statistic lies in the extreme 5% of the fitted gamma. If the statistic lies in the bottom 2.5% then a blue (light) shading is made in a clockwise direction indicating the proportion of bottom tail that has been undercut. The shading for Θ_2 indicate that the diagnostic was just in the bottom tail, suggesting that beliefs about Θ_2 have changed slightly less than anticipated *a priori*. Red (dark) diagnostic shadings are drawn in an anti-clockwise direction, and are used to highlight statistics in the upper tail, similarly. None are present in the example figures.

This graph is useful for highlighting the overall picture after introduction of $\{X_1, \ldots, X_6\}$ into the graph. After six observations, it can be seen that Θ_5 is the state variable about which most has been learned. Also, slightly more is learned about the observation residuals, ν_t, than the state residuals, ω_t. Diagnostically, adjustments related to time points 3 and 4 led to changes in beliefs smaller than anticipated *a priori*, but otherwise no problems are highlighted.

Figure 2 shows the effect of sequentially introducing X_1 to X_6 in a stepwise manner. These diagrams are intended to be colour, allowing effects of observation of a particular node to be matched up easily with its effect on other nodes of the moral graph — the colour of the central portion of each observed node is the colour used to for the portions removed from other nodes on the graph. Looking again at Θ_2, it can be seen that most of the uncertainty resolved for this node was due to the first three observations, and that the third observation contributed least of the three. However, from a diagnostic viewpoint, there are no shadings corresponding to the first two observations, but severe blue (light, clockwise) shadings corresponding to the information from the third observation. This indicates that beliefs about Θ_2 changed much less than anticipated due to the introduction of the third observation into the graph. It is presumably the third observation which contributed most to the small diagnostic shading on the global picture.

There are several things worth noting about this graph. Clearly, for any particular time, t, no observations from before time t are informative for the residuals (ν_t and ω_t) at time t. Also, whilst only the current observation provides (non-negligible) information for ω_t, the current and next observation are informative for ν_t. This is why analysing estimated residuals from a sequentially applied Kalman filter (without back-forecasting) is misleading, as there, the estimated residuals for time t are based on observations up to time t only. Note also that information about the state flows forwards and backwards in time, some considerable distance. The diagnostic shadings are all blue (light, clockwise), corresponding to changes in belief smaller than expected, but for the completely observed nodes, only the observed value of X_4 was particularly surprising.

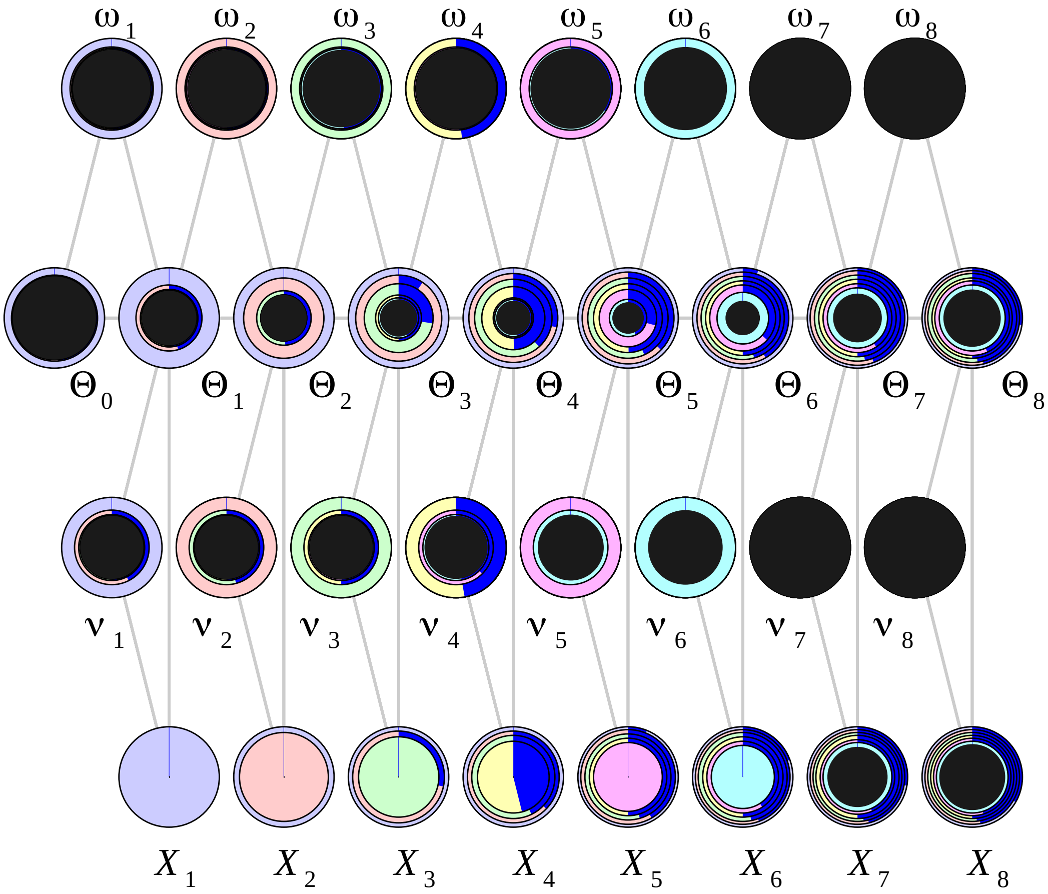

Figure 2: Stepwise moral graph

Figure 3 shows a snapshot of the junction tree, illustrating the partial effect of the introduction of the sixth observation. Looking at the arc between A_4 and B_4, the (very light) outer portion indicates the maximum amount of information which can flow down this particular arc. The inner-shaded portion shows proportion of this that can flow down the arc due to the observation of X_6. Half of the arc is shaded blue (light), indicating that the message that actually flowed down this arc was weaker than anticipated (according to the same diagnostic rule used for the moral graph figures). This picture is mainly of interest when there are diagnostic problems on the moral graph, and one is interested in pinning down their origin. This graph is intended to give some insight into the way computations are being carried out at the clique-tree level. Some blue (light) shadings indicate changes in belief smaller than expected with respect to the last partial adjustment (that is, with respect to the adjusted structure after five observations have already been made), perhaps indicating messages weakening faster than anticipated

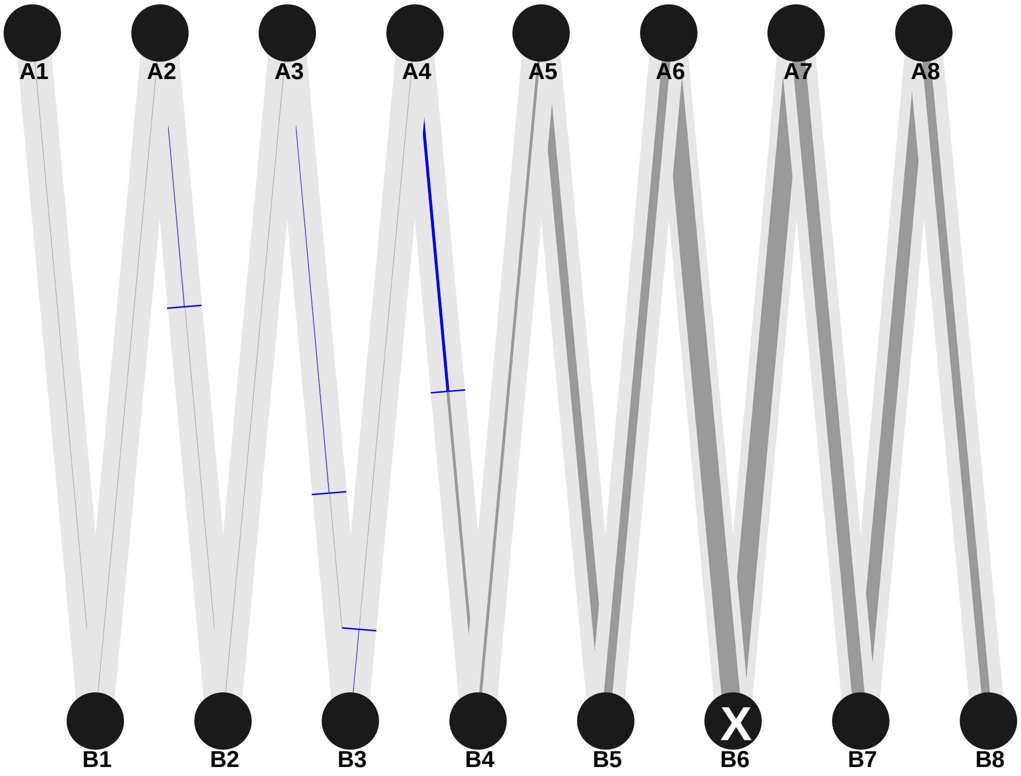

Figure 3: Junction tree snapshot graph

as they flow back in time, giving rise to the blue shadings in Figure 2. The fixed width of the outer portions of the arcs indicate that a similar amount of information can (potentially) flow down each arc. The inner portions of the arcs get slimmer the further one gets from the point that information was introduced, showing the diminishing effect of the information on far-away nodes.

5 Example: Local computation for matrix objects

In order to illustrate the generality of the geometric approach to local computation here presented, the algorithms will be applied to the problem of local computation in a Hilbert space of random matrix objects.

The previous example required specification to be made for two noise covariance matrices (V and W), for which no updating occurs in the simple Bayes linear framework. Let us now suppose that we wish to update our beliefs about these matrix objects, based on observable data. Using the approach adopted in Wilkinson & Goldstein (1997), the V and W matrices will be treated as objects within an inner-product space of random matrices, and beliefs will be adjusted by orthogonally projecting into a space of observable random matrices.

Construction of the space of random matrices begins with a basis for the space of constant $r \times r$ matrices. Define $C_{ij} = e_i e_j^T$ (e_i is an r-dimensional vector with a 1 in the ith position, and zeros elsewhere), and let $C = \{C_{ij} | 1 \leq i \leq r, 1 \leq j \leq r\}$. Also define collections of observable random matrices

$$\tilde{D}_n = \{X_3^{(1)} X_3^{(1)^T}, X_4^{(1)} X_4^{(1)^T}, \ldots, X_n^{(1)} X_n^{(1)^T}, X_3^{(2)} X_3^{(2)^T}, X_4^{(2)} X_4^{(2)^T}, \ldots, X_n^{(2)} X_n^{(2)^T}\}, \quad \forall n > 3 \tag{39}$$

where

$$X_t^{(j)} = X_t - X_{t-j} \tag{40}$$

Let $D_n = C \cup \tilde{D}_n$, $\forall n > 3$ and $B = D_\infty$. Then $\langle B \rangle$ is a real vector space. Impose the inner-product

$$(F, G) = \mathrm{E}(\mathrm{trace}(F^T G)), \quad \forall F, G \in \langle B \rangle \tag{41}$$

on $\langle B \rangle$, and complete the resulting inner-product space into a Hilbert space, $[B]$. Note that $[B]$ will contain many important limit points not found in $\langle B \rangle$. In particular, under the assumption of exchangeability of the quadratic products of residuals, we may use the second-order exchangeability representation theorem to decompose uncertainty about these in the following way.

$$\nu_t \nu_t^T = V^\nu + S_t^\nu \tag{42}$$

$$\omega_t \omega_t^T = V^\omega + S_t^\omega \tag{43}$$

Note that V^ν and V^ω are matrices representing the uncertain covariance matrices underlying the DLM, and in particular, that $\mathrm{E}(V^\nu) = V$ and $\mathrm{E}(V^\omega) = W$. V^ν and V^ω are limit points in the space $[B]$.

Belief adjustment will be carried out in this space by forming $\mathrm{E}_{D_n}(V^\nu)$ and $\mathrm{E}_{D_n}(V^\omega)$, the orthogonal projections of V^ν and V^ω into $[D_n]$, respectively. Note that the projection of a matrix into the constant space, $[C]$, gives the expected value of the matrix, *ie.* $\mathrm{E}(X) = \mathrm{E}_C(X)$. Belief separation in $[B]$ takes the form

$$(X \perp\!\!\!\perp Z)/Y \iff (X - \mathrm{E}_Y(X), Z - \mathrm{E}_Y(Z)) = 0, \; X, Y, Z \in [B] \tag{44}$$

and so graphical models can be formed and manipulated, and local computation may be carried out, in a similar way to that already outlined.

Now, for scalar adjustments, the inner-product $(X, Y) = \mathrm{E}(XY)$ is used, but it is the covariance matrices of the "zero-adjusted" inner-products, $\langle X, Y \rangle = (X - \mathrm{E}(X), Y - \mathrm{E}(Y))$ which are propagated around the tree, and used for adjustment purposes. Similarly for adjustments in this matrix space, it is the matrices of the $\langle \cdot, \cdot \rangle$ inner-products which are used and propagated.

Due to the n-step exchangeability properties of the observable random matrices (Wilkinson & Goldstein 1997) the moral graph for the observable matrices together with V^ν and V^ω has

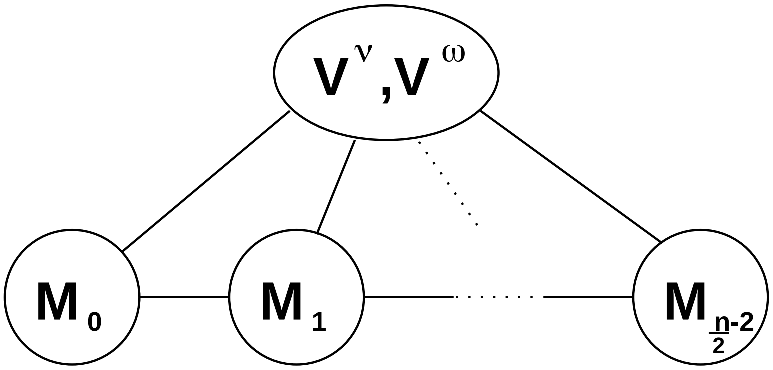

Figure 4: Moral graph for the matrix objects

a sparse, but rather complex structure, which lives most naturally in 4-dimensional space. However, by carefully grouping together collections of matrices, a simple 2-dimensional moral graph can be obtained (in the sense that it can easily be projected into 2-dimensional space without lots of crossed arcs). The structure of this graph is shown in Figure 4, where

$$M_i = \left\{ X^{(1)}_{2i+3} {X^{(1)}_{2i+3}}^T, X^{(1)}_{2i+4} {X^{(1)}_{2i+4}}^T, X^{(2)}_{2i+3} {X^{(2)}_{2i+3}}^T, X^{(2)}_{2i+4} {X^{(2)}_{2i+4}}^T \right\}, \quad \forall i \geq 0 \tag{45}$$

The cliques of the moral graph take the form

$$C_i = \left\{ \{V^\nu, V^\omega\} \cup M_{i-1} \cup M_i \right\}, \quad \forall i \geq 1 \tag{46}$$

and form a simple Markov chain on which we may carry out local computation. Suppose we are interested in design questions; in particular, how many observations we would need in order to reduce uncertainty about the covariance matrices to an acceptable level. We are now in a position to answer such questions.

Using the inner-product specifications from Wilkinson & Goldstein (1997), the clique tree is initialised as follows.

$$\langle C_i, C_i \rangle = \begin{pmatrix} 10592 & 0 & 21183 & 21183 & 21183 & 21183 & 21183 & 21183 & 21183 & 21183 \\ 0 & 6894 & 6894 & 6894 & 13788 & 13788 & 6894 & 6894 & 13788 & 13788 \\ 21183 & 6894 & 2311526 & 327822 & 951204 & 951204 & 49260 & 49260 & 334716 & 56154 \\ 21183 & 6894 & 327822 & 2311526 & 334716 & 951204 & 327822 & 49260 & 951204 & 334716 \\ 21183 & 13788 & 951204 & 334716 & 3797012 & 244974 & 56154 & 56154 & 348504 & 69942 \\ 21183 & 13788 & 951204 & 951204 & 244974 & 3797012 & 334716 & 56154 & 244974 & 348504 \\ 21183 & 6894 & 49260 & 327822 & 56154 & 334716 & 2311526 & 327822 & 951204 & 951204 \\ 21183 & 6894 & 49260 & 49260 & 56154 & 56154 & 327822 & 2311526 & 334716 & 951204 \\ 21183 & 13788 & 334716 & 951204 & 348504 & 244974 & 951204 & 334716 & 3797012 & 244974 \\ 21183 & 13788 & 56154 & 334716 & 69942 & 348504 & 951204 & 951204 & 244974 & 3797012 \end{pmatrix}$$

Observations are introduced into the moral graph in a sequential fashion, and the effects on the "uncertainty" associated with $\{V^\nu, V^\omega\}$ are shown in Figure 5, where the shading conventions used are as for Figure 2, though obviously without diagnostic shadings, as a preposterior analysis is being carried out here. The large node represents the pair of uncertain covariance matrices, $\{V^\nu, V^\omega\}$. The figure shows the effect of sequential observations of 14, 28, and 40 time points of observable matrix objects. The dark central portion represents the uncertainty remaining after 40 observations. As preposterior analysis is straightforward within the Bayes linear approach, this local computation theory is ideal for tackling complex design questions for large statistical models. In addition, the geometric formulation allows these techniques to apply to general collections of random entities — not just random scalars.

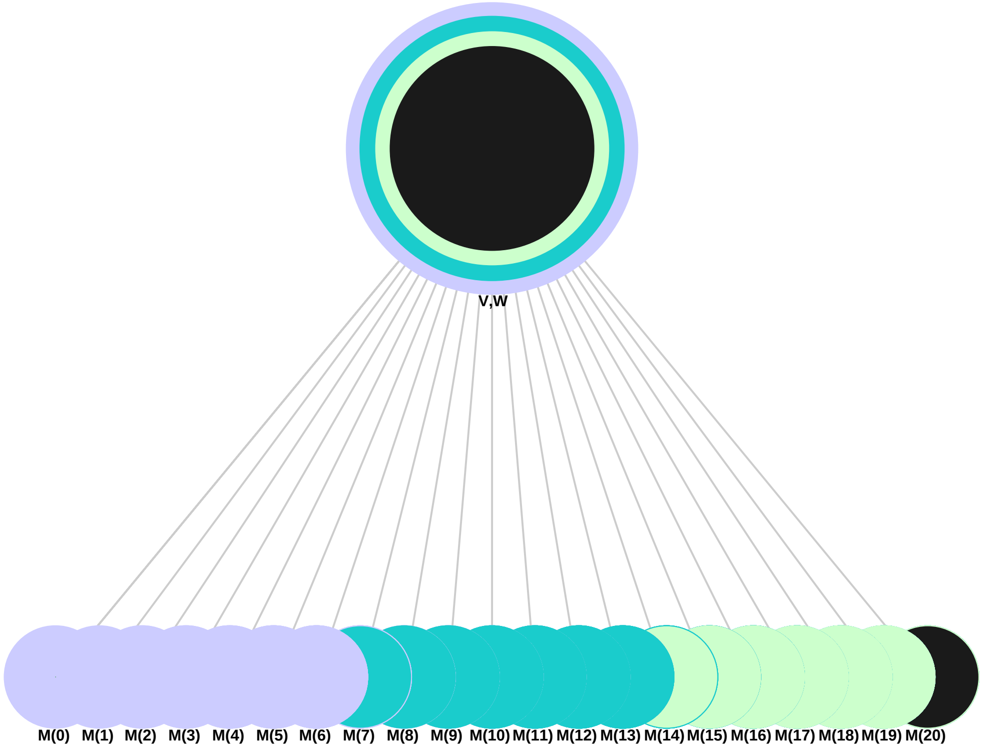

Figure 5: Uncertainty resolution for the matrix objects

6 Implementation of the examples

Both of the examples in this paper were implemented using the freely available software, BAYES-LIN (Wilkinson 1997). This software is an extension of the LISP-STAT object-oriented environment for statistical computing (Tierney 1990). The software provides a collection of object prototypes and associated methods appropriate for carrying out local computation *via* pure message-passing between objects representing the junction-tree nodes of Bayes linear belief networks.

The software defines high-level primitives for the definition and initialisation of objects representing junction tree nodes. Methods are associated with these objects which provide a full mechanism for local computation. Data is introduced into a particular object by sending it an appropriate message. The object uses the information to calculate the covariance between the data and itself, then passes this matrix on to neighbouring junction tree objects. On receiving such a message from another object, the receiving object uses the matrix to calculate the covariance between the data and itself, using the separating set between them, before then passing a similar message on to its neighbours. In this way, the covariance matrix between the data and the nodes represented by the objects are propagated around the junction tree. Adjustments and absorptions are computed in a similar way. The system also provides primitives for the construction of moral graph nodes corresponding to the junction tree, and allows direct introduction of information into this graph, which is then passed on to a junction tree node that contains it, for propagation around the tree. Graphical object prototypes are also defined, which allow interactive production of full colour versions of the interpretive and diagnostic figures described in this paper.

An example file is provided with the BAYES-LIN package which demonstrates the construction of and local computation for a multivariate DLM similar to the one discussed in this paper.

7 Conclusions

7.1 The Bayes linear approach to local computation

Here we have presented a complete geometric framework for local computation of belief adjustments, and the accompanying interpretive and diagnostic features, for Bayes linear belief networks. In addition, a freely available set of programming tools for carrying out local computation within such a framework has been briefly described. The theory and algorithms open up the possibility of using Bayes linear methods to tackle inferential problems at least an order of magnitude more complex than previously possible. Furthermore, a set of interpretive and diagnostic graphical displays have been proposed which allow a vast amount of summary information relating to the adjustment process to be more easily comprehended.

The general advantages of the Bayes linear approach to graphical models are as follows. Firstly, we are allowed complete flexibility for the prior specifications, without the pretence that we can make full probabilistic specifications over such complex structures. Secondly, since all computations reduce to matrix algebra, the algorithms remain tractable even for problems where the cliques become very large (several hundred variables per clique). Thirdly, since we have taken a general geometric approach to the problem, the methodology extends straightforwardly to allow inference in more general spaces, such as spaces of covariance matrices. Fourthly, since adjusted variance is not a function of the observed data, the methodology allows full preposterior analysis to take place, so that local computation for complex design problems becomes tractable within this framework. Finally, our methodology may be viewed as a pragmatic approximation to a full Bayes analysis when such an approach would be too complex. However, usually we may only make specifications over limited aspects of our prior beliefs, so it is more honest to develop methods which only require partial prior specifications. The foundational justifications for the methodology presented here may be found in Goldstein (1997), which is concerned with temporal inference under partial prior specification.

7.2 Comparison with other methodologies

Pearl (1988) provides a comprehensive overview of local computation for Bayesian networks, although most of the algorithms developed apply only to networks with probabilistic nodes. Section 7.2 of that volume is concerned with networks for continuous variables, but the algorithm developed applies only to directed tree graphs with univariate nodes. Normand & Trichler (1992) extend this algorithm to directed trees with multivariate nodes, but still do not allow general directed acyclic graphs with multivariate nodes. Lauritzen (1992) provides just such a generalisation for mixed models based on Conditional Gaussian (CG) distributions. Applying his algorithm to a continuous variable network gives identical updates to those obtained using the procedures outlined here, but the algorithm presented here is much cleaner, as it does not require any formal construction of separators. Furthermore, Lauritzen's algorithm is highly dependent on extremely restrictive distributional assumptions, and can only be used for sequential updating if one is prepared to adopt (non-Bayesian) *ad hoc* re-normalisation procedures after each update. Of course, the distributional assumptions are made in order to reduce the updating procedure to a tractable set of linear matrix equations. In contrast, we do not make any distributional assumptions, but rely on exactly the same linearity properties of adjustments in order to arrive at a formally identical analysis; we believe that this is a more honest approach to belief updating for many complex inferential problems.

In addition to the points already made, it is clear that all of the above methods apply solely to updating of random (scalar) quantities, and fail to emphasise the geometric nature of the construction. On the contrary, we have emphasised the purely geometric nature of belief separation, and a geometric framework for local computation. Thus, our methodology readily extends to inference in general spaces of random entities; in particular, we have demonstrated application to inference in spaces of random matrices. Furthermore, no other approach has been developed to the extent that a full set of interpretive and diagnostic summaries of the adjustment process is locally computed for graphical display.

References

Dawid, A. P. (1979), 'Conditional independence in statistical theory', *J. Roy. Statist. Soc.* **B:41,1**, 1–31.

Dawid, A. P. & Lauritzen, S. L. (1993), 'Hyper Markov laws in the statistical analysis of decomposable graphical models', *Ann. Statist.* **21**, 1272–1317.

de Finetti, B. (1974), *Theory of probability, vol. 1*, Wiley.

Goldstein, M. (1981), 'Revising previsions: a geometric interpretation', *J. R. Statist. Soc.* **B:43**, 105–130.

Goldstein, M. (1988), The data trajectory, *in* J.-M. Bernardo et al., eds, 'Bayesian Statistics 3', Oxford University Press, pp. 189–209.

Goldstein, M. (1990), Influence and belief adjustment, *in* J. Smith & R. Oliver, eds, 'Influence Diagrams, Belief Nets and Decision Analysis', Wiley, Chichester.

Goldstein, M. (1997), Prior inferences for posterior judgements, *in* M. L. D. Chiara et al., eds, 'Structures and norms in science', Pordrecht Kluwer.

Goldstein, M. (1998), Bayes linear analysis, *in* 'Encyclopedia of Statistical Sciences'.

Goldstein, M., Farrow, M. & Spiropoulos, T. (1993), 'Prediction under the influence: Bayes linear influence diagrams for prediction in a large brewery', *The Statistician* **42**(2), 445–459.

Goldstein, M. & Wooff, D. A. (1995), 'Bayes linear computation: concepts, implementation and programming environment', *Statistics and Computing* **5**, 327–341.

Jensen, F. V. (1996), *An introduction to Bayesian networks*, UCL Press.

Lauritzen, S. L. (1992), 'Propagation of probabilities, means, and variances in mixed graphical association models', *J. Amer. Statist. Assoc.* **87,420**, 1098–1108.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. & Leimer, H. G. (1990), 'Independence properties of directed markov fields', *Networks* **20**, 491–505.

Normand, S.-L. & Trichler, D. (1992), 'Parameter updating in a Bayes network', *J. Amer. Statist. Assoc.* **87,420**, 1109–1115.

Pearl, J. (1988), *Probabilistic reasoning in intelligent systems*, Morgan Kaufmann.

Smith, J. Q. (1990), Statistical principles on graphs, *in* J. Smith & R. Oliver, eds, 'Influence Diagrams, Belief Nets and Decision Analysis', Wiley, Chichester.

Tierney, L. (1990), *LISP-STAT: An object oriented environment for statistical computing and dynamic graphics*, Wiley.

Wilkinson, D. J. (1997), 'BAYES-LIN: An object-oriented environment for bayes linear local computation', `http://www.ncl.ac.uk/~ndjw1/bayeslin/`.

Wilkinson, D. J. & Goldstein, M. (1996), Bayes linear adjustment for variance matrices, *in* J.-M. Bernardo et al., eds, 'Bayesian Statistics 5', University Press, Oxford, pp. 791–800.

Wilkinson, D. J. & Goldstein, M. (1997), 'Bayes linear covariance matrix adjustment for multivariate dynamic linear models', in submission.