LA-UR-16-27699
Approved for public release; distribution is unlimited.

Title: Doubly-Hierarchical One-Way Random Effects Model: Multivariate Data

Author(s): Sigeti, David Edward
Vander Wiel, Scott Alan

Intended for: Report

Issued: 2016-10-06

Doubly-Hierarchical One-Way Random Effects Model: Multivariate Data

David E. Sigeti, LANL, XCP-8, sigeti@lanl.gov
Scott Vander Wiel, LANL, CCS-6, scottv@lanl.gov

October 4, 2016

Contents

1 Introduction and the Likelihood

This report presents the mathematical definition of a doubly-hierarchical one-way random effects model for multivariate data. Multivariate data with m components y_i $(i = 1, \ldots, n)$ arise from G groups with the vector $\widehat{g}[i] \in \{1, \ldots, G\}$ denoting the group of the i-th observation. Data within

each group are modeled as independent samples from a multivariate normal distribution specific to that group. Thus, the likelihood for the multivariate datum y_i is,

$$\left[y_i \mid \theta_{\widehat{g}[i]}, \Sigma_{\widehat{g}[i]}\right] \overset{\text{ind}}{\sim} \text{Normal}\left(\theta_{\widehat{g}[i]}, \Sigma_{\widehat{g}[i]}\right), \tag{1}$$

where, θ_g and Σ_g are, respectively, the unknown mean vector and covariance for the g-th group[1].

2 The Hierarchical Prior for the Group Means

If we were simply doing Bayesian inference on each group separately, then we would pick priors for the mean vectors and covariance matrices of each group[2] and use Bayes law to obtain posterior distributions for the means and standard deviations. Such priors could potentially be different for each group but, in the simplest case, we could pick common priors for all groups. For example, we could pick a normal prior for the group means,

$$\left[\theta_g \mid \mu_\theta, \Sigma_\theta\right] \overset{\text{ind}}{\sim} \text{Normal}\left(\mu_\theta, \Sigma_\theta\right), \tag{2}$$

where μ_θ and Σ_θ are, respectively, the mean vector and the covariance matrix for the prior for the group means.

In such a separate analysis, μ_θ and Σ_θ are fixed and therefore known inputs to the analysis. However, if we believe that the distributions of the separate groups should be similar to each other and that the data from one group should therefore provide us with information about other groups (especially groups for which we do not yet have any data), then we can treat the inputs to the above common prior for the group means as themselves unknown, and use the data to make inferences about these unknown "hyperparameters"[3].

Because μ_θ and Σ_θ are now unknown and are to be inferred from the data, we need priors for them. We choose a non-informative, improper prior for the location hyperparameter μ_θ,

$$\mu_\theta \propto \mathbf{1}_m\big([-\infty, \infty]\big), \tag{3}$$

[1]Note that, in this and all later equations specifying the model, all quantities are assumed statistically independent except as explicitly stated conditionally.

[2]Or, more generally, we would pick a joint prior for the mean and covariance of each group.

[3]Note that another way to come to the same approach is to assume that the separate group means, θ_g, are themselves independently sampled from the normal distribution in Equation 2.

where $\mathbf{1}_m\big([a,b]\big)$ is the function that is equal to one on the m-dimensional hypercube defined by vectors a of lower limits and b of upper limits and zero elsewhere.

In order to find an appropriate prior for Σ_θ, we follow the recommendations in [5, Section 6.12] and begin by decomposing it into a correlation matrix and a diagonal matrix of variances,

$$\Sigma_\theta = \text{diag}\left(\sqrt{T_\theta}\right) C_\theta \, \text{diag}\left(\sqrt{T_\theta}\right) \tag{4}$$

where:

- T_θ is the diagonal of Σ_θ (in other words, the vector of variances);
- We can take the square root of T_θ because it contains only nonnegative elements;
- The operator diag converts a vector to a diagonal matrix; and
- C_θ is the correlation matrix corresponding to Σ_θ.

We must now choose priors for T_θ and C_θ. We begin by choosing proper priors for the components of T_θ that we will subsequently make quite broad and, thus, "weakly informative" [2, Section 2.9]

$$\left[T_{\theta_j} \mid \alpha_{\theta_j}, \beta_{\theta_j}\right] \sim \Gamma^{-1}\left(\alpha_{\theta_j}, \beta_{\theta_j}\right), \tag{5}$$

where,

- Γ^{-1} is the inverse-gamma distribution, the conjugate prior for the normal variance;
- α_θ and β_θ are vectors of length m that specify the standard shape and scale parameters, respectively, for the inverse-gamma distributions that are the priors for the components of T_θ;

Here, α_θ and β_θ are fixed and known inputs to the analysis. We will choose them based on our experience with choosing parameters for the inverse-gamma prior for the variance in the univariate case.

As in the univariate case, we choose,

$$\alpha_{\theta_j} \equiv 0.3, \ \forall j. \tag{6}$$

This specifies a very wide prior with a very long tail as the variance goes to infinity.

As for the scale β_θ, we want to choose it based on a (vague) prior estimate of a typical value of the variance in the group means. However, we cannot simply set β_θ to such a scale because, as discussed at greater length in [4], the parameter β in the inverse-gamma distribution is not a "typical" value from the distribution because reasonable measures of central tendency like the mean, median, and mode all depend strongly on α as well as β. In addition, neither the mean nor the mode of the distribution are always appropriate because the mean does not even exist for $\alpha \leq 1$, and the mode goes to a fixed value of β as $\alpha \to 0$, even though the distribution is becoming more and more weighted toward ∞ without limit. In practice then, the only reasonable choice for a typical value from the inverse-gamma distribution is the median because it always exists and scales reasonably with α over the entire range from $0 \to \infty$.

There is no analytic formula for the median of the inverse-gamma distribution, but it is shown in [4] that $\gamma(\alpha, \beta)$, defined by,

$$\gamma\left(\alpha, \beta\right) \;=\; \beta \, \frac{\sqrt[\epsilon]{\left(Ae^{\ln(2)/\alpha} - A\right)^{\epsilon} + 1}}{\sqrt[\epsilon]{\alpha^{\epsilon} + 1}}, \quad A \equiv 1.7996, \quad \epsilon \equiv 1.15,$$

approximates the median to within 4% for $\alpha \in [10^{-3}, 10^{3}]$, which is more than good enough for our purposes.

We can therefore invert the definition of γ to obtain $\beta(\alpha, \gamma)$,

$$\beta\left(\alpha, \gamma\right) \;=\; \gamma \, \frac{\sqrt[\epsilon]{\alpha^{\epsilon} + 1}}{\sqrt[\epsilon]{\left(Ae^{\ln(2)/\alpha} - A\right)^{\epsilon} + 1}}, \quad A \equiv 1.7996, \quad \epsilon \equiv 1.15, \tag{7}$$

and choose γ to be any reasonable prior scale for the variance.

In the case at hand, we therefore use Equation 7 to set

$$\beta_{\theta_j} = \beta\left(\alpha_{\theta_j}, S_{\theta_j}\right) \tag{8}$$

where S_{θ_j} is a vector of scales for the variances of the group means, one for each component of the multivariate data, and is another fixed input to the analysis.

We also have to specify a prior for the correlation matrix C_θ. We choose to use the LKJ prior for correlation matrices discussed in [2, Appendix A],

$$p\left(C_\theta \,|\, \eta_\theta\right) \equiv \text{LKJCorr}\left(C_\theta \,|\, \eta_\theta\right) \propto \left|C_\theta\right|^{\eta_\theta - 1}, \quad \eta_\theta > 0, \tag{9}$$

where η_θ is another fixed input to the analysis. For $\eta_\theta = 1$, the univariate marginal densities in the LKJCorr prior for the correlations are uniform on

$[-1, 1]$. For $\eta_\theta > 1$, the univariate marginals go to zero at -1 and 1, while for $\eta_\theta < 1$, they are infinite at 1 and -1. We will start by setting

$$\eta_\theta = 1 \tag{10}$$

and change to a somewhat larger value if the MCMC has problems with C_θ becoming numerically singular (as it will if any of the correlations get too close to ± 1).

3 The Hierarchical Prior for the Group Variances

We next need to specify a statistical model for the Σ_g, the group covariance matrices in Equation 1. We choose a common, inverse-Wishart model for the Σ_g,

$$[\Sigma_g \mid \nu, R_v] \overset{\text{ind}}{\sim} W^{-1}(\nu, R_v), \tag{11}$$

where ν is a scalar degrees-of-freedom parameter and R_v is a symmetric positive-definite scale matrix. We then assume that ν and R_v are unknown hyperparameters and therefore are to be inferred from the data.[4]

As was the case with our hierarchical model for the group means, in order to infer our hyperparameters (in this case ν and R_v) from the data, we need (hyper)priors for them. We begin with ν. In Section A.3 we note that a diagonal element, $\Sigma_{v_{ii}}$, of a covariance matrix Σ_v distributed according to Equation 11 will have an inverse-gamma marginal distribution,

$$\Sigma_{v_{ii}} \sim \Gamma^{-1}(\alpha, \beta), \quad \alpha \equiv (\nu - k + 1)/2, \quad \beta \equiv R_{v_{ii}}/2. \tag{12}$$

where $R_{v_{ii}}$ is the corresponding diagonal element of the scale matrix, R_v. We therefore choose a minimum value for ν based on the minimum value for α that we used in the univariate case [6],

$$\nu_{\min} \equiv 2\alpha_{\min} + k - 1. \tag{13}$$

We then set the prior for ν analogously to the prior for α in the univariate case,

$$[\nu \mid \nu_{\min}, S_\nu] \sim \text{Normal}_{+1/2}(\nu_{\min}, S_\nu), \tag{14}$$

[4] Equivalently, we can regard the Σ_g as sampled independently from the prior in Equation 11.

where the subscript $+1/2$ on a distribution indicates the half-distribution to the right of the median, and $\nu_{\min}$ and S_ν represent a minimum value and scale for ν, respectively. We will set these quantities based on our previous experience with the univariate model through the connection to the α-parameter for the inverse-gamma distributions for the variances given by Equation 12,

$$\begin{aligned} \nu_{\min} &\equiv 2\alpha_{\min} + k - 1, \\ S_\nu &\equiv 2S_\alpha + k - 1, \end{aligned} \tag{15}$$

where $\alpha_{\min}$ and S_α correspond to the quantities used to specify the hyperprior for α in the univariate case. We will choose

$$\alpha_{\min} \equiv 0.2 \text{ and } S_\alpha \equiv 100, \tag{16}$$

at least initially, providing a broad hyperprior for ν, with the underlying α allowed to take values between 0.2 and approximately 200. We do this with the understanding that we may need to raise $\alpha_{\min}$ or lower S_α if we experience problems with the MCMC.

As for the hyperprior for the symmetric positive-definite scale matrix R_v, we will follow the same strategy that we used for the covariance matrix for the group means and will specify separate priors for the diagonal elements on the one hand and the corresponding correlation matrix on the other. We write,

$$R_v = \text{diag}\left(\sqrt{T_v}\right)\, C_v\, \text{diag}\left(\sqrt{T_v}\right), \tag{17}$$

where T_v is the vector of diagonal elements of R_v and C_v is the correlation matrix corresponding to R_v.

Based on Equation 12, a diagonal element of R_v, $R_{v_{ii}}$, corresponds to $2\beta_{v_i}$ where β_{v_i} is the scale parameter of the inverse-gamma marginal distribution for the corresponding diagonal element of a covariance matrix sampled from the inverse-Wishart distribution[5]. We follow the same strategy that we used earlier and define a vector of parameters $\gamma_v \in \mathcal{R}^m$ which are approximations to the medians for the inverse-gamma distributions for each variance. We choose a hyperprior for γ_{v_i},

$$\left[\gamma_{v_i} \mid S_{v_i}\right] \;\sim\; \text{Cauchy}_{+1/2}\left(0.0, S_{v_i}\right), \tag{18}$$

[5]Note that all of inverse-gamma distributions for the diagonal elements of the sampled covariance matrices will have the same value for the parameter α, namely $\alpha = (\nu+k-1)/2$, so there is not need for a vector of αs.

where S_{v_i} is an independent input to the analysis that represents a rough scale for the group variance of the i-th component.

We can then use Equation 7 to compute,

$$\beta_{\nu_i} = \beta(\alpha, \gamma_{\nu_i}), \tag{19}$$

and compute the actual diagonal element of R_v, T_{v_i} from Equation 12,

$$T_{v_i} \equiv 2\beta_{v_i}. \tag{20}$$

The quantity α used to compute β_{v_i} in Equation 7 is, of course, derived from the parameter ν according to Equation 12,

$$\alpha = (\nu - k + 1)/2. \tag{21}$$

We also have to specify a prior for the correlation matrix C_v. As in Equation 9, we use the LKJ prior for correlation matrices,

$$p\left(C_v\right) \equiv \text{LKJCorr}(C_v \,|\, \eta_v) \propto |C_v|^{\eta_v - 1}, \quad \eta_v > 0, \tag{22}$$

where η_v is another fixed input to the analysis. As above, we will start out trying

$$\eta_v = 1, \tag{23}$$

and change to a somewhat larger value if the MCMC has problems with C_v becoming singular.

4 Scale Inputs

We are thus left with the following inputs that we need to specify for the analysis,

1. S_θ, a vector of prior estimates for the scales of the variances of the group means;
2. S_v, a vector of prior estimates for the scales of the within-group variances.

We choose to depart from strict Bayesianism and choose both of these inputs to be equal to the sample variance of the corresponding component of the multivariate data over the entire dataset,

$$S_{\theta_i} \equiv S_{v_i} \equiv \text{ the variance of the } i\text{-th component over the dataset.} \tag{24}$$

Given that the priors that S_θ and S_v specify are weakly informative, this choice simply insures a reasonable scale for the our hyperpriors.

5 Parameters, Inputs, and Construction of the Posterior Distribution

We have now provided a complete description of our doubly-hierarchical model. The unknown quantities that we need to infer are,

$$\begin{aligned} \theta_g,\, \Sigma_g : &\quad \text{the parameters for each group,} \\ \mu_\theta,\, T_\theta,\, C_\theta : &\quad \text{the hyperparameters for the prior for the group means,} \\ \nu,\, \gamma_v,\, C_v : &\quad \text{the hyperparameters for the prior for the group variances,} \end{aligned} \tag{25}$$

where $g \in \{1, \ldots, G\}$

We construct a posterior for these unknowns in accordance with Bayes law by multiplying the likelihoods for the data in Equation 1 by the prior for the group means in Equation 2 and the hyperpriors in Equations 5 and 9 for the hyperparameters for the prior for the group means, using Equation 4 to connect the hyperparameters T_θ and C_θ to the covariance matrix Σ_θ in the prior in Equation 2.

We then multiply by the prior for the group covariance matrices, Equation 11, and then by the hyperpriors for the hyperparameters in the prior for the group covariances in Equations 14, 18, and 22, using Equations 19, 20 and 21 to compute T_v from ν and γ_v and Equation 17 to assemble the scale matrix for the inverse-Wishart prior for the group variances from T_v and C_v.

The fixed inputs to our analysis are,

$$\begin{aligned} \alpha_{\theta_j} : &\quad \text{(see Equation 6),} \\ S_{\theta_i} : &\quad \text{(see Equation 24),} \\ \eta_\theta : &\quad \text{(see Equation 10),} \\ \alpha_{\min} : &\quad \text{(see Equation 16),} \\ \alpha_{\text{scale}} : &\quad \text{(see Equation 16),} \\ S_{v_i} : &\quad \text{(see Equation 24),} \\ \eta_v : &\quad \text{(see Equation 23),} \end{aligned}$$

A Appendix: The Inverse-Wishart Distribution

In this section, we present some facts about the inverse-Wishart distribution that we use in the body of the report.

We begin immediately below by defining and relating the two conventions for defining the inverse-Wishart distribution. We then present an important theorem for marginal distributions for parts of a partitioned matrix distributed according to an inverse-Wishart, and then specialize the theorem to give the marginal distribution for the variances that are the diagonal elements of a random matrix from an inverse-Wishart.

A.1 Conventions for the Inverse-Wishart Distribution

There appear to be two distinct conventions for defining the inverse-Wishart distribution that differ in the specification of the degrees-of-freedom parameter. In this subsection, we show the correspondence between the two conventions.

The first convention appears in the text by Muirhead, where we find [3, problem 3.6, page 113]:

> A random $m \times m$ positive definite matrix B is said to have the inverted Wishart distribution with n degrees of freedom and positive definite $m \times m$ parameter matrix V if its density function is,
>
> $$\frac{2^{-m(n-m-1)/2}}{\Gamma_m\left[\frac{1}{2}(n-m-1)\right]} \frac{(\det V)^{(n-m-1)/2}}{(\det B)^{n/2}} \operatorname{etr}\left(-\tfrac{1}{2}B^{-1}V\right), \; B > 0, \tag{26}$$
>
> where $n > 2m$. We will write that B is $W_m^{-1}(n, V)$.

Note that, for a matrix A, $\operatorname{etr}(A)$ is $\exp[\operatorname{trace}(A)]$ and Γ_m is the multivariate gamma function.

Wikipedia gives what appears to be the current standard convention for the inverse-Wishart distribution for a symmetric positive-definite $p \times p$ matrix X [7]

$$W^{-1}(X \mid \Psi, \nu) \equiv \frac{|\Psi|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)} |X|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\operatorname{tr}\left(\Psi X^{-1}\right)}, \tag{27}$$

where $\nu > p - 1$ and Ψ is a symmetric positive-definite $p \times p$ scale matrix.

If we make the obvious equivalences,

$$\begin{aligned} m &= p \\ B &= X \\ V &= \Psi, \end{aligned} \tag{28}$$

and remember that matrices commute within a trace, then the two definitions are equivalent if,

$$\nu = n - m - 1 \tag{29}$$

or, equivalently

$$n = \nu + m + 1. \tag{30}$$

As a check, it is trivial to show that the two conditions on the degrees-of-freedom parameter are equivalent,

$$\nu > p - 1 \;\Leftrightarrow\; n - m - 1 > m - 1 \;\Leftrightarrow\; n > 2m.$$

A.2 Marginal Distribution for a Partitioned Covariance

Given the definition in Equation 26 above (see also [1]), Muirhead states the following [3, problem 3.6(d), page 114]:

> Suppose that B is $W_m^{-1}(n, V)$ and partition B and V as
>
> $$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \quad V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix},$$
>
> where B_{11} and V_{11} are $k \times k$ and B_{22} and V_{22} are $(m-k)\times(m-k)$. Show that B_{11} is $W_k^{-1}(n - 2m + 2k, V_{11})$.

According to Wikipedia [7], if $A \in R^{p\times p}$ is a symmetric nonnegative-definite matrix with an inverse-Wishart distribution, $W^{-1}(\Psi, \nu)$, and A and Ψ are partitioned conformably,

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad \Psi = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix}, \tag{31}$$

with $A_{11}, S_{11} \in R^{p_1\times p_1}$, $A_{22}, S_{22} \in R^{p_2\times p_2}$, where clearly $p_1 + p_2 = p$, then,

$$A_{11} \;\sim\; W^{-1}(\Psi_{11}, \nu - p_2)\,. \tag{32}$$

We will now check that the two formulas are equivalent. We begin by noting the following equivalences (on top of the equivalences in Equations 28, 29, and 30),

$$\begin{aligned} A_{ij} &= B_{ij}, \\ p_1 &= k, \\ p_2 &= m - k = p - p_1. \end{aligned} \tag{33}$$

We will define,

$$\begin{aligned} n' &\equiv n - 2m + 2k \\ \nu' &\equiv \nu - p_2. \end{aligned} \tag{34}$$

We want to show that Equation 30 applies to the new pair of matrices, B_{11} and A_{11}. In other words, we want to show that,

$$n' = \nu' + k + 1,$$

where the m on the right hand side of Equation 30 is now a k because A_{11} and B_{11} are $k \times k$ matrices. If we then plug the definitions of n' and ν' from Equation 34 into Equation 30, we have,

$$\begin{aligned} n - 2m + 2k &= (\nu - p_2) + p_1 + 1 \\ &= (\nu - (m - k)) + k + 1 \\ &= \nu - m + k + k + 1 \end{aligned}$$

which is obviously equivalent to

$$n - m = \nu + 1,$$

and thus to

$$n = \nu + m + 1,$$

so the degrees-of-freedom parameters for B_{11} and A_{11} obey Equation 30 if and only if the original parameters for A and B do.

A.3 Marginal Distributions for Variances (Diagonal Elements)

Adopting the notation from Wikipedia [7], if we set $p_1 \equiv 1$ in Equation 32, then we have for the first diagonal element of A,

$$A_{11} \;\sim\; W^{-1}(\Psi_{11}, \nu - p - 1)\,. \tag{35}$$

Furthermore, again according to [7], the one-dimensional inverse-Wishart distribution, $W_1^{-1}(\Psi, \nu)$, where Ψ is now a positive scalar, is equivalent to the inverse-gamma distribution with $\alpha \equiv \nu/2$ and $\beta \equiv \Psi/2$. Thus,

$$A_{11} \;\sim\; \Gamma^{-1}(\alpha, \beta), \quad \alpha \equiv (\nu - k + 1)/2, \quad \beta \equiv \Psi_{11}/2. \tag{36}$$

References

[1] Taras Bodnar and Yarema Okhrin, *Properties of the singular, inverse and generalized inverse partitioned wishart distributions*, Journal of Multivariate Analysis **99** (2008), 2389–2405.

[2] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, *Bayesian data analysis*, third ed., Chapman and Hall/CRC, 2013.

[3] R. J. Muirhead, *Aspects of multivariate statistical theory*, Wiley, New York, 1982.

[4] David E. Sigeti, *Notes on the inverse-gamma distribution*, Tech. Report LA-UR-15-27468, Los Alamos National Laboratory, Los Alamos, NM, 2015.

[5] Stan Development Team, *Stan modeling language: Users guide and reference manual*, mc-stan.org, December 2015, Stan Version 2.9.0.

[6] Scott Vander Wiel and David E. Sigeti, *Doubly-hierarchical one-way random effects model*, Tech. Report LA-UR-15-27467, Los Alamos National Laboratory, Los Alamos, NM, 2015.

[7] Wikipedia, *Inverse-Wishart distribution — Wikipedia, The Free Encyclopedia*, 2016, [Online; accessed 8-March-2016].