Bayesian learning in hierarchical joint models

Hatfield, Laura (1st author)
University of Minnesota, Division of Biostatistics
A460 Mayo, 420 Delaware St SE
Minneapolis, MN 55454, USA
E-mail: hatfield@umn.edu

Hodges, James (2nd author)
University of Minnesota, Division of Biostatistics
2221 University Ave SE, Suite 200
Minneapolis, MN 55414, USA
E-mail: hodges@ccbr.umn.edu

Carlin, Brad (3rd author)
University of Minnesota, Division of Biostatistics
A460 Mayo, 420 Delaware St SE
Minneapolis, MN 55454, USA
E-mail: brad@biostat.umn.edu

Introduction

Investigators often collect multiple types of data to learn about an underlying biological process. For example, in clinical trials, the effects of treatment on disease may be quantified by observing both the longitudinal progression of symptoms and the survival time until a clinical event. Analysis of these outcomes separately may ignore important dependencies, such as the informative censoring of further symptom data by the event. Therefore, a broad class of joint models for longitudinal and time-to-event data has been developed. The most flexible such models rely on an underlying latent variable framework to induce dependence among outcome types. The development of these approaches has greatly expanded the scope of models to accommodate many data complexities, see Tsiatis and Davidian (2004) for a review. Yet little attention has been paid to their properties and performance. A few notable exceptions include work on residuals (Rizopoulos et al., 2010) and specific cases in which joint modeling may not be justified (Hanson et al., 2011). Our contribution to this literature concerns the relative contributions of the two classes of data to inference. In particular, we wish to understand how longitudinal and survival data contribute to learning about model parameters of greatest scientific interest.

Previous authors (Reich and Hodges, 2008) have explored reparameterizing linear mixed models to elucidate the roles of data and random effect variance parameters. Their work depends crucially on use of a normal-normal model for clever manipulations of the matrix expressions to yield separation of the posterior into interpretable components (Hodges, 1998). We extend that line of inquiry to the joint modeling setting in which latent effects link a longitudinal submodel to a survival submodel.

For well-behaved posterior distributions (i.e., unimodal and approximately symmetric), information content may be reasonably quantified by inverse variance (aka precision). When the prior and posterior are both proper and available in closed form, precision is accessible, and we attempt to partition this information measure into portions due to the longitudinal and survival data, respectively. The normal-normal model we consider here is simple enough to allow this approach. However, for joint models of greater complexity, the posteriors are *not* generally available analytically, and analysis must proceed by Markov chain Monte Carlo (MCMC).

First, we motivate the problem using a few empirical results from a complex real data set and

joint model. Next, we derive expressions posterior variances in a simplified joint model, considering both latent effects and fixed effects of interest.

Model Notation

Let the vector of longitudinal observations for individual i be denoted $\mathbf{y}_i$, the survival outcome be t_i with censoring indicator $\delta_i = 1$ if the event is observed, and $\mathbf{u}_i$ be a vector of person-specific latent variables. We adopt the usual joint modeling assumption that the $\mathbf{u}_i$ induce all of the associations between $\mathbf{y}_i$ and t_i, so that they are conditionally independent given $\mathbf{u}_i$. Writing the rest of the parameters as $\mathbf{\Theta}$, the specification of a joint distribution for longitudinal and survival outcomes is simplified since we can write $f(\mathbf{y}_i, t_i|\mathbf{u}_i, \mathbf{\Theta}) = f(\mathbf{y}_i|\mathbf{u}_i, \mathbf{\Theta})f(t_i|\mathbf{u}_i, \mathbf{\Theta})$. Collecting the longitudinal, survival, and latent variables across subjects into $\mathbf{Y}$, $\mathbf{T}$, and $\mathbf{U}$, respectively, we write the likelihood as $f(\mathbf{Y}, \mathbf{T}|\mathbf{U}, \mathbf{\Theta}) = f(\mathbf{Y}|\mathbf{U}, \mathbf{\Theta})f(\mathbf{T}|\mathbf{U}, \mathbf{\Theta})$. The joint posterior distribution is $f(\mathbf{U}, \mathbf{\Theta}|\mathbf{Y}, \mathbf{T}) = \frac{f(\mathbf{Y},\mathbf{T}|\mathbf{\Theta},\mathbf{U})\pi(\mathbf{\Theta},\mathbf{U})}{\int f(\mathbf{Y},\mathbf{T}|\mathbf{\Theta},\mathbf{U})\pi(\mathbf{\Theta},\mathbf{U})d(\mathbf{\Theta},\mathbf{U})}$ and we focus on various marginal posteriors derived from this. In particular, we are interested in treatment effects in $\mathbf{\Theta}$ and individual-level parameters in $\mathbf{U}$, where we wish to quantify the relative contribution of the two data types to the posterior inferences on these. For example, suppose $\boldsymbol{\beta}_{12}$ is an element of $\mathbf{\Theta}$ that quantifies the effect of treatment on longitudinal trajectories. We wish to study how its posterior depends on features of the data and model such as the error variance of both data types, the amount of censoring in the survival outcomes, and the distribution of the latent variables.

To write the specific Normal-Normal model that we will use below, denote the longitudinal observations for the i^{th} subject by $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$, where y_{ij} is observed at time s_{ij}, $j = 1, \ldots, n_i$. Assume that the longitudinal observations are normal $y_{ij} \sim N(\mathbf{x}_{1ij}\boldsymbol{\beta}_1 + u_i, \sigma_1^2)$, the log survival times arise from a normal distribution $\log(t_i) \sim N(\mathbf{x}_{2i}\boldsymbol{\beta}_2 + \alpha u_i, \sigma_2^2)$, and the scalar latent effects u_i are independently and identically $N(0, \sigma_u^2)$. The design vectors $\mathbf{x}_{1ij}$ and $\mathbf{x}_{2i}$ contain covariates of interest, including treatment effects, and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ parameterize their effects on the mean of the longitudinal variables and the median of the log survival times, respectively. We use prior distributions that are proper, vague, and conjugate if possible.

Results from a Clinical Trial Setting

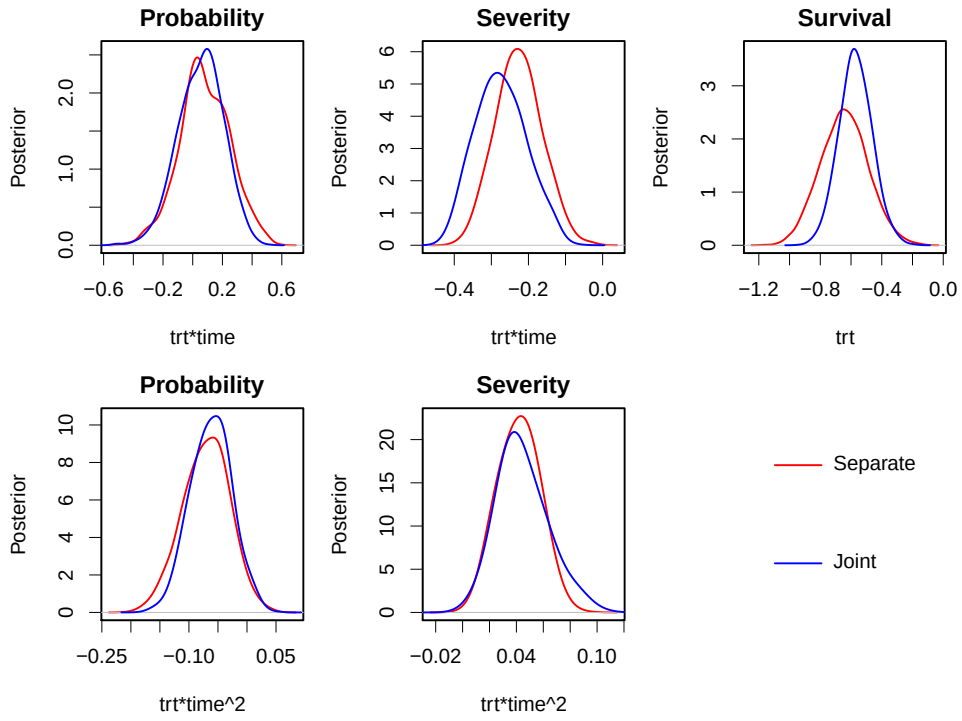
Our clinical trial data derive from a large Phase III clinical trial of first-line therapy for malignant pleural mesothelioma, a rapidly fatal lung disease. Participants were randomized to one of two treatment conditions and self-reported their symptom presence and severity weekly during the treatment phase. They were then followed until substantial progression of the disease or death (whichever occurred first) to determine progression-free survival time. Clinical interest focused on the effects of treatment on the longitudinal trajectories of symptom presence and severity and on PFS.

Consider the zero-augmented beta (ZAB) distribution developed in previous work (Hatfield et al., 2011a,b), which characterizes observations on the support $Y \in \{0\} \cup (0, 1)$ using a three-parameter distribution having $\omega = Pr(Y \in (0, 1))$, $\mu = E(Y|Y \in (0, 1))$, and ϕ a dispersion parameter. We build regression submodels for ω and μ. In particular, the longitudinal observation y_{ij} for the i^{th} person's j^{th} observation (at time s_{ij}) is governed by $\text{logit}(\omega_{ij}) = \mathbf{x}_{0ij}\boldsymbol{\beta}_0 + u_{0i}$ and $\text{logit}(\mu_{ij}) = \mathbf{x}_{1ij}\boldsymbol{\beta}_1 + u_{1i1} + u_{1i2}s_{ij}$. The design vectors are $\mathbf{x}_{0ij} = \mathbf{x}_{1ij} = (1, s_{ij}, s_{ij}^2, trt_i * s_{ij}, trt_i * s_{ij}^2)$, that is, we assume the time trend is quadratic and the treatment effects are multiplicative, since participants were randomized to treatment and there are no baseline differences between the two groups. Then we specify a Weibull distribution for the survival model, which is a two-parameter distribution having hazard parameter λ and shape parameter γ. We build a regression model for the hazard parameter $\log(\lambda_i) = \mathbf{x}_{2i}\boldsymbol{\beta}_2 + \alpha_1 u_{0i} + \alpha_2 u_{1i1} + \alpha_3 u_{1i2}$. The design vector is $\mathbf{x}_{2i} = (1, trt_i)$ and we take a simple

linear combination of the person-specific parameters. These latent variables are assumed to come from a zero-centered multivariate normal distribution with 3×3 covariance matrix Σ_u.

A simple way of approaching the question of attributing information to the two data types for a given model and dataset is to fit the joint model and compare the resulting posteriors to those from reduced models that utilize only longitudinal or survival data, respectively. To study the influence of survival data on the longitudinal treatment effect elements of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$, we compare their posteriors estimated from this joint model to one with all the same structure except no Weibull submodel. To study the influence of longitudinal data on the survival treatment effect element of $\boldsymbol{\beta}_2$, we must modify the approach. The survival submodel specified above cannot be fit on its own with the linear combination of three person-specific effects and only one survival observation per person. Thus, we modify the approach to use a single random effect, $\log(\lambda_i) = \mathbf{x}_{2i}\boldsymbol{\beta}_2 + u_i$, which is then scaled by α_0 and α_1 for addition to the linear predictors of the two longitudinal submodels, respectively. Then we can fit the survival submodel alone or in the context of the joint model and examine the differences.

Figure 1: Smoothed posterior density estimates for treatment effects in models fit using the longitudinal and survival data separately or jointly.



The results of these comparisons are shown in Figure 1. This displays the posterior distribution for the five parameters that represent treatment effects: the linear and quadratic treatment-by-time interactions for symptom presence (left column), the linear and quadratic treatment-by-time interactions for symptom severity (middle column), and the treatment effect on the hazard of progression/death (right column). Notice that the addition of survival data impacts the treatment effects in the longitudinal submodels very little (left four panels) compared to the impact of adding longitudinal data to the esimation of treatment effects in the survival submodel (top right panel). This seems to indicate that the longitudinal data are more informative in this model.

Learning About Latent Effects

We turn next to the derivation of analytical expressions in a simplified joint model. In the following, will use time-independent regression models having only intercepts and treatment effects for both the longitudinal and survival submodels, that that is, $\mathbf{x}_{1ij} = \mathbf{x}_{1i} = (1, trt_i)'$ and $\mathbf{x}_{2i} = (1, trt_i)'$, where trt_i is an indicator of the treatment group of the i^{th} individual. Using to the simple normal-normal joint model formulation, $y_{ij} \sim N(\beta_{11}+\beta_{12}trt_i+u_i, \sigma_1^2)$ and $\log(t_i) \sim N(\beta_{21}+\beta_{22}trt_i+\alpha u_i, \sigma_2^2)$. Notice that by removing the time-dependence of the mean of the longitudinal observations, we can write $\bar{y}_{i\cdot} \sim N(\beta_{11} + \beta_{12}trt_i + u_i, \sigma_1^2/n_i)$. For notational simplicity, we define $\bar{y}_{i\cdot} = z_{1i}$, $\log(t_i) = z_{2i}$, and $\sigma_1^2/n_i = \sigma_{1i}^2$.

Focusing on the posterior for an individual's latent parameter u_i suppose the survival time is observed. The conditional mode of u_i (conditional on the data z_{1i}, z_{2i} and the remaining parameters $\boldsymbol{\Theta}$) is

$$u_i^* = \left(\frac{z_{1i} - \beta_{11} - \beta_{12}trt_i}{\sigma_{1i}^2} + \frac{\alpha(z_{21} - \beta_{21} - \beta_{22}trt_i)}{\sigma_2^2} \right) \sigma_{u_i}^2 \ ,$$

where $\sigma_{u_i}^2 = \left(\frac{1}{\sigma_{1i}^2} + \frac{\alpha^2}{\sigma_2^2} + \frac{1}{\sigma_u^2} \right)^{-1}$. Notice that the posterior mode for uncensored survival observations is increasing with the sum of scaled residuals of the linear predictors from the longitudinal and survival submodels. The second derivative of the log posterior is negative everywhere, $\frac{\partial^2 \log f(u_i|\boldsymbol{\Theta}, z_{1i}, z_{2i})}{\partial u_i^2} = -\sigma_{u_i}^{-2}$ and the Fisher information is summed precisions, also sensible. The posterior variance of u_i shrinks with greater precision in either of the longitudinal or survival data.

Learning About Fixed Effects

Following previous work (Reich and Hodges, 2008) showing that posteriors depend on the ratio of sampling to random effect variance, we consider looking for similar expressions in the simple Normal-Normal model. Consider the longitudinal treatment effect β_{12}, then we are interested obtaining the posterior of β_{12} and studying how it depends on $\mathbf{z}_1$, $\mathbf{z}_2$, and the variances σ_1^2, σ_2^2, and σ_u^2.

We recall the hierarchical modeling result of Lindley and Smith (1972). For $n \times p_1$ response vector $\mathbf{Y}$, p_1-vector of parameters θ_1, known $n \times p_1$ design matrix A_1, and known $n \times n$ covariance matrix C_1, let the likelihood be $\mathbf{Z} \sim N(A_1\boldsymbol{\theta}_1, C_1)$. Then for second-level p_2-vector of parameters θ_2, known design and covariance matrices A_2 and C_2, let the prior be $\theta_1 \sim N(A_2\boldsymbol{\theta}_2, C_2)$. The authors showed that the marginal distribution is $\mathbf{Z} \sim N(A_1A_2\boldsymbol{\theta}_2, C_1 + A_1C_2A_1')$ and the posterior is $\boldsymbol{\theta}_1|\mathbf{z} \sim N(D\mathbf{d}, D)$ where $D^{-1} = A_1'C_1^{-1}A_1 + C_2^{-1}$ and $\mathbf{d} = A_1'C_1^{-1}\mathbf{z} + C_2^{-1}A_2\boldsymbol{\theta}_2$. Because we assume all the covariance parameters (including α) are known in the first two cases above, we can apply these results.

To simplify the expressions, we make the following assumptions without loss of generality: $n_i = n \ , \forall i$ (thus $\sigma_{1i}^2 = \sigma_1^2/n$) there are equal numbers of subjects in treatment and control groups (i.e., $M = M^T$), and the trt_i covariate is coded so that $trt_i = 1$ indicates observations from the treatment group and $trt_i = -1$ from the control group. Then we collect the data into a single $2N$-vector, where longitudinal outcomes come first, sorted into treatment then control outcomes, followed by survival outcomes, similarly sorted: $\mathbf{Z} = (z_{11}, \ldots, z_{1,N/2}, z_{1,N/2+1}, \ldots, z_{1N}, z_{21}, \ldots, z_{2,N/2}, z_{2,N/2+1}, \ldots, z_{2N})'$. The complete $(4 + N)$-vector of parameters contains both fixed and latent effects $\boldsymbol{\theta}_1 = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \mathbf{u})'$. The covariance matrix C_1 is block diagonal since conditional on u_i all the responses are independent; the upper left block for the longitudinal observations is $\sigma_1^2/n\mathbf{I}_N$ and the lower right block for the survival observations is $\sigma_2^2\mathbf{I}_N$. We place independent normal priors on $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)'$ with mean $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)'$ and with variances $\sigma_{\beta_1}^2$ and $\sigma_{\beta_2}^2$, respectively. We use another independent normal prior distribution on $\mathbf{u}$, centered at $\mathbf{0}$ with common variance σ_u^2. Then the joint posterior covariance matrix on $(\boldsymbol{\beta}, \mathbf{u})'$ can be partitioned into a 4×4 submatrix V_11 governing the fixed effects, an $N \times N$ submatrix V_22 for the random effects, and off-diagonal submatrices containing the covariances between these two types of parameters. The properties of multivariate normal distributions make it simple to obtain the *marginal* posterior covariances by taking corresponding diagonal elements of this matrix. Thus we

find the posterior variance of β_{12} to be

$$(1) \quad \frac{P_2 - c_4^2 N P_u^{-1}}{P_1 P_2 - N P_u^{-1}\left[c_4^2 P_1 + c_3^2 P_2\right]}$$

where $P_1 = \left(\frac{Nn}{\sigma_1^2} + \frac{1}{\sigma_{\beta_1}^2}\right)$, $P_2 = \left(\frac{N}{\sigma_2^2} + \frac{1}{\sigma_{\beta_2}^2}\right)$, $P_u = \left(\frac{n}{\sigma_1^2} + \frac{\alpha^2}{\sigma_2^2} + \frac{1}{\sigma_u^2}\right)$, $c_3 = \frac{n}{\sigma_1^2}$, and $c_4 = \frac{\alpha}{\sigma_2^2}$. If we assume improper flat priors on the βs so that $\frac{1}{\sigma_{\beta_2}^2}, \frac{1}{\sigma_{\beta_1}^2} \to 0$, the variance of β_{12} simplifies to $\frac{n\sigma_u^2 + \sigma_1^2}{Nn}$, which surprisingly does not involve σ_2^2 at all.

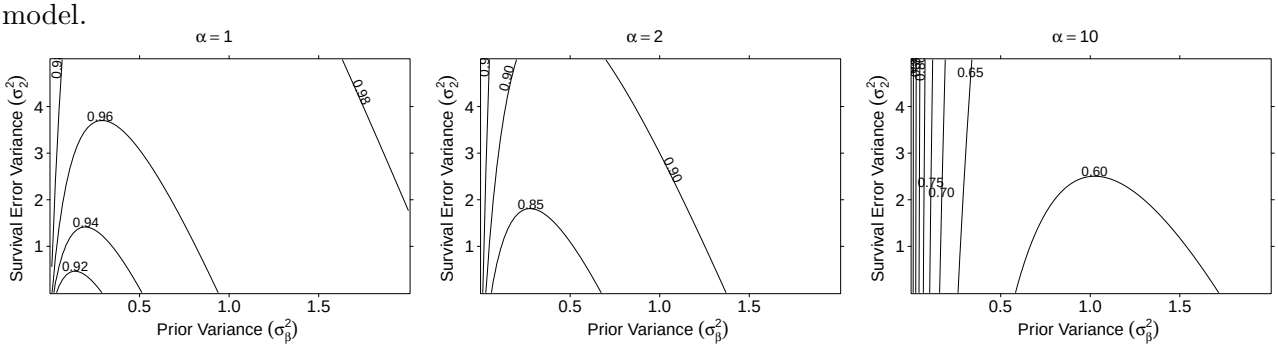
We compare the expression for β_{12}'s posterior variance in the joint model to the case of fitting a longitudinal-data-only model of the same form. Again following the L&S notation, the elements A_1, C_1, $\boldsymbol{\theta}_1$, and C_2 are simply the reduced forms obtained by deleting the longitudinal data and parameters. Following the same procedure above, we obtain the posterior variance of β_{12}

$$\left[P_1 - N c_3^2 \left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_u^2}\right)^{-1}\right]^{-1}.$$

However, we see that we can obtain the same result by setting $\alpha = 0$ in (1) above, since P_u becomes $\left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_u^2}\right)$ and P_2 cancels out of the remaining terms.

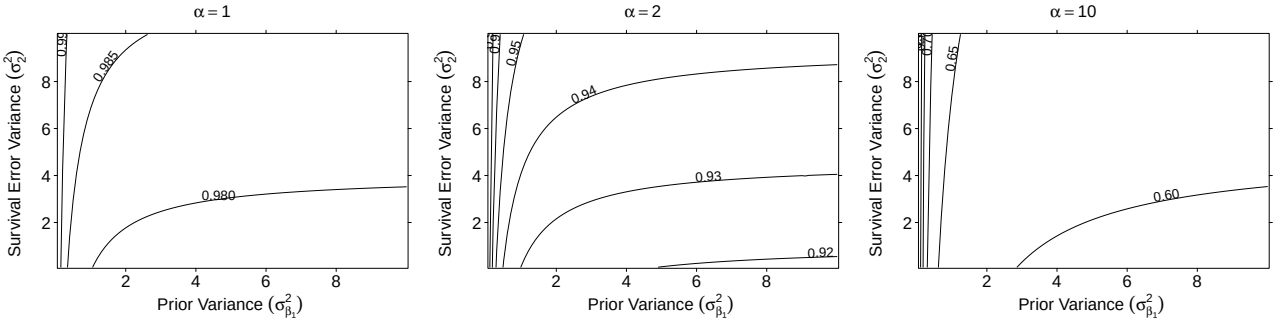
Then we consider the variance reduction achieved by adding survival data to the longitudinal data in this setting, i.e., the ratio of $Var(\beta_{12}|\alpha, \mathbf{Z})$ to (1) to $Var(\beta_{12}|\alpha = 0, \mathbf{Z})$. Figure 2 displays this ratio as it depends on σ_2^2 and $\sigma_{\beta_1}^2 = \sigma_{\beta_2}^2 = \sigma_\beta^2$ for three values of α, where we have fixed the remaining conditioning variables ($\sigma_1^2 = 1$, $n = 2$, $N = 10$). In this figure, small values of the ratio represent the most benefit of adding survival data to the posterior variance of the longitudinal treatment effect, since it will be small compared to the case of no survival information. In the left panel ($\alpha = 1$), the dependence of the longitudinal treatment effect variance on survival error variance is minimal; at the minimum, the ratio is only 0.9. In the right panel ($\alpha = 10$), more benefit is gained by including the survival data– the minimum ratio is 0.6. We can we see that the minimum variance is achieved when $\sigma_2^2 = 0$, which is sensible since this represents survival data that perfectly determine each u_i. The joint minimum is achieved when the prior variance is neither too large nor too small, $\sigma_\beta^2 = \alpha/N$. At this value of the prior variance, the dependence on σ_2^2 is greatest (i.e., the gradient in the direction of σ_2^2 is steepest).

Figure 2: Ratio of posterior variance of longitudinal treatment effect in joint model to $\alpha = 0$ separate model.



However, we may not wish to constrain the prior variances in both parts of the model to be equal. Figure 3 displays the same ratio where $\sigma_{\beta_2}^2 = 2$ is fixed and $\sigma_{\beta_1}^2$ varies along the x axis. In this figure, we clearly see that larger values of α lead to smaller ratios (i.e., greater benefit of adding survival) and that the gradient in the direction of the survival error variance (y direction in this figure) depends on the variance of the prior on β_{12}. Clearly the minimum in σ_2^2 will again be at 0, but the variance does not reach a unique minimum in $(\sigma_{\beta_1}^2, \sigma_2^2)$ jointly.

Figure 3: Ratio of posterior variance of longitudinal treatment effect in joint model to $\alpha = 0$ separate model.



Conclusion

We have derived some expressions for the contributions of the two types of data to the posteriors of model parameters of interest in joint models. These show that the complex interactions of multiple variance parameters– including those for longitudinal and survival errors, latent parameters, and priors– interact in complex ways to determine the posterior variances of treatment effects.
The modeling simplifications required to make analytical progress have taken the models far from the complex joint models frequently seen in the literature. Therefore, future work will seek to extend this partitioning of information to more general modeling settings using simulation approaches.

REFERENCES (RÉFERENCES)

Chen, M.-H. and Gustafson, P. (eds.) (2011), *Lifetime Data Analysis*, Springer, vol 17, Special Issue.

Hanson, T., Branscum, A., and Johnson, W. (2011), "Predictive comparison of joint longitudinal-survival modeling: a case study illustrating competing approaches," *Lifetime Data Analysis*, 17, 3–28.

Hatfield, L., Boye, M., and Carlin, B. (2011a), "Joint modeling of multiple longitudinal patient-reported outcomes and survival," *Journal of Biopharmaceutical Statistics*, (To Appear).

Hatfield, L., Hackshaw, M., Boye, M., and Carlin, B. (2011b), "Bayesian joint models of zero-inflated longitudinal patient-reported outcomes and progression-free survival times in mesothelioma." (Unpublished).

Hodges, J. (1998), "Some algebra and geometry for hierarchical models, applied to diagnostics," *Journal of the Royal Statistical Society, Series B*, 60, 497–536.

Lindley, D. and Smith, A. (1972), "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society, Series B*, 34, 1–41.

Reich, B. and Hodges, J. (2008), "Identification of the variance components in the general two-variance linear model," *Journal of Statistical Planning and Inference*, 138, 1592–1604.

Rizopoulos, D. (2011), "Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data," *Biometrics*, (To Appear).

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2010), "Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes," *Biometrics*, 66, 20–29.

Schoop, R., Schumacher, M., and Graf, E. (2011), "Measures of prediction error for survival data with longitudinal covariates," *Biometrical Journal*, 53, 275–293.

Tsiatis, A. and Davidian, M. (2004), "Joint modeling of longitudinal and time-to-event data: an overview," *Statistica Sinica*, 14, 809–834.