Published in final edited form as:
Comput Stat Data Anal. 2012 June 1; 56(6): 1303–1318. doi:10.1016/j.csda.2011.09.004.

Frailty Modeling via the Empirical Bayes Hastings Sampler

Richard A. Levine[a,*], **Juanjuan Fan**[a], **Pamela Ohman Strickland**[b], and **Shaban Demirel**[c]

Richard A. Levine: ralevine@sciences.sdsu.edu; Juanjuan Fan: jjfan@sciences.sdsu.edu; Pamela Ohman Strickland: ohmanpa@umdnj.edu; Shaban Demirel: sdemirel@deverseye.org

[a]Department of Mathematics and Statistics, 5500 Campanile Drive, San Diego State University, San Diego, CA, 92182

[b]Division of Biometrics, UMDNJ, Piscataway, NJ 08854

[c]Devers Eye Institute, Discoveries in Sight Research, Laboratories, 1225 NE 2nd Ave, Portland, OR 97232

Abstract

Studies of ocular disease and analyses of time to disease onset are complicated by the correlation expected between the two eyes from a single patient. We overcome these statistical modeling challenges through a nonparametric Bayesian frailty model. While this model suggests itself as a natural one for such complex data structures, model fitting routines become overwhelmingly complicated and computationally intensive given the nonparametric form assumed for the frailty distribution and baseline hazard function. We consider empirical Bayesian methods to alleviate these difficulties through a routine that iterates between frequentist, data-driven estimation of the cumulative baseline hazard and Markov chain Monte Carlo estimation of the frailty and regression coefficients. We show both in theory and through simulation that this approach yields consistent estimators of the parameters of interest. We then apply the method to the short-wave automated perimetry (SWAP) data set to study risk factors of glaucomatous visual field deficits.

Keywords

multivariate survival analysis; nonparametric Pólya tree prior; Gibbs sampler; Metropolis-Hastings sampler; goodness of fit; glaucoma and ophthalmology data

1 Introduction

Analyses of data from studies of visual field deficits and glaucomatous progression are complicated by correlations between observed failure times from fellow eyes of a subject. Bayesian frailty models have proven to be a valuable tool for modeling this dependence through a random effect term in a proportional hazards model. However, the practitioner is left to choose from a wide array of frailty distributions, the choice of which may affect inferences drawn on parameters (hazard ratios) of interest. Not to mention, the dependence structure is unknown presenting difficulties in parameterizing a frailty model and exposing "default" models, such as a gamma frailty distribution, as seemingly arbitrary.

[*]Corresponding author: Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Drive, San Diego, CA, 92182; 619-594-6494; ralevine@sciences.sdsu.edu.

A nonparametric approach to frailty modeling provides a flexible alternative in which the frailty distribution is left unspecified, letting the data a posteriori drive the functional form. In such models both the frailty distribution and baseline hazard rate are modeled nonparametrically. The nonparametric frailty term presents no difficulties in the construction of a Markov chain Monte Carlo (MCMC) algorithm for drawing posterior inferences. Standard Gibbs samplers for fitting nonparametric Bayesian models (e.g., Walker and Mallick, 1997) may be applied for sampling full conditional distributions on the frailties and the parametric portion of the proportional hazards model. However, incorporation of the baseline hazard into the Markov chain Monte Carlo (MCMC) routine turns out to be a challenging task. A wide array of models for the baseline hazard and MCMC methods for fitting these models have been proposed in the literature (for example, see Ibrahim et al., 2001). However, as Gustafson et al. (2003) mentions in the motivation of their work, the routines are computationally and mathematically intensive and not easily automated, leaving the non-expert with a difficult task in applying such inferential procedures.

In our motivating application the primary goal is to infer hazard rates, studying risk factors for glaucoma and short wavelength automated perimetry for detecting visual field defects. The baseline hazard rate is then effectively a nuisance parameter (function). An inferential routine for the baseline hazard which requires complicated mathematical derivation and substantial computational coding and implementation cost is clearly undesirable. In this paper we propose an empirical Bayes approach to alleviate difficulties in modeling the baseline hazard and subsequently incorporating it into an MCMC algorithm. The idea derives from the work of Casella (2001) in which we use the data to “estimate away” nuisance parameters, focusing computational and inferential effort on the parameters of interest. In the nonparametric Bayesian frailty model, we estimate the baseline hazard through a nonparametric frequentist estimator and then construct an MCMC algorithm to iteratively simulate posterior samples conditional on this empirical Bayes estimate. The method draws on the deep theory and vast implementation options of MCMC and EM algorithms. Furthermore, the routine is simple to automate within the construct of the Gibbs and Hastings samplers for fitting nonparametric Bayesian models. We argue that this empirical Bayes Hastings sampler requires less coding time and computational expense than the popular piecewise hazard approaches (e.g., Walker and Mallick, 1997), requires less tweaking of tuning and model parameters in fact lending to complete automation.

The Bayesian frailty model with nonparametric specification of the frailty distribution is best suited for our study of glaucomatous progression. However, the empirical Bayes Hastings sampler in this setting, as a general approach, lends to diagnostic tools for testing parametric forms for the frailty distribution and routines for performing model selection. We highlight these issues in our analysis of glaucomatous visual field defects. Furthermore, the proposed modeling and inferential strategies provide a flexible framework within which to mix and match nonparametric and parametric components and strategies for handling nuisance parameters.

In Section 2, we formally define the nonparametric Bayesian frailty model, expressing the frailty distribution nonparametrically through a Pólya tree process. We also define the nonparametric estimator of the cumulative baseline hazard to be incorporated into our empirical Bayes routine. In Section 3 we, primarily for notational purposes, briefly detail the Pólya tree distribution. In Section 4, we introduce the empirical Bayes Hastings sampler for drawing inferences under the semi-parametric frailty model, estimating the baseline hazard rate in a Monte Carlo E-type step in the MCMC routine. As part of the discussion of the Hastings sampler, we derive conditions under which the random variates drawn reasonably represent a sample from the posterior distribution of interest. We also discuss issues for optimally implementing the MCMC sampling scheme in practice. Section 5 presents

simulation studies to validate our proposed methods for drawing inferences under the frailty model. In Sections 6 and 7, we present routines for computing Bayes factors and traverse the regression parameter space to evaluate parametric forms of the frailty distribution and perform variable selection within our empirical Bayes Hastings sampler framework. In Section 8, we illustrate our proposed methods in the analysis of a data set for studying glaucomatous visual field deficits. In Section 9 we conclude with a discussion of practical issues beyond the developments and applications in this paper.

2 Frailty Model

Suppose that the observed data consist of clustered, and possibly censored, failure-time data represented by $Y_{ik} = \{X_{ik}, \delta_{ik}, Z_{ik}\}$, with $k = 1, \cdots, K_i$ and $i = 1, \cdots, n$. $X_{ik} = T_{ik} \wedge C_{ik}$ is the minimum of the failure time and the censoring time; $\delta_{ik} = I\{(X_{ik} = T_{ik})\}$, the failure indicator, which takes the value of 1 if $(X_{ik} = T_{ik})$ and 0 otherwise; and Z_{ik} is a p-vector of covariates. It is assumed that the failure time vector $T_i = (T_{i1}, \cdots, T_{iK_i})'$ is independent of the censoring time vector $C_i = (C_{i1}, \cdots, C_{iK_i})'$ given $Z_i = (Z_{i1}', \cdots, Z_{iK_i}')'$, $i = 1, \cdots, n$.

The proportional hazards model (Cox, 1972) has been widely applied in analyzing independent or univariate failure times. As a generalization of the Cox proportional hazards model for clustered or multivariate failure times, Clayton and Cuzick (1985) introduce the frailty model in which a random effect term (or "frailty") is assumed to have a multiplicative effect on the hazard. In terms of the hazard function, the model can be stated as follows:

$$\lambda_{ik}(t|\boldsymbol{Z}_{ik}, V_i) = \lambda_0(t)\exp(\beta' \boldsymbol{Z}_{ik})V_i \quad (1)$$

where $\lambda_0(t)$ is an unknown baseline hazard function, $\boldsymbol{\beta}$ is a p-vector of unknown regression parameters, and V_i is the frailty, representing some common unobserved characteristics shared by all the failure times in the ith cluster. It is assumed that, given the frailty V_i, failure times within the ith cluster are independent. Note that the baseline hazard function λ_0 may be assumed to depend on k, for example, in a family study when $k = 1, 2$ refers to mothers and daughters, respectively.

Let $\theta_i = \ln V_i$ for each $i = 1, \ldots, n$ with the n-vector of log-frailties denoted by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$. We will avoid difficulties in specifying the frailty distribution by modeling this distribution nonparametrically. In particular, following Walker and Mallick (1997), assume

$$\begin{aligned} \theta_1, \ldots, \theta_n \quad & iid \quad F \\ F \quad & \sim \quad PT(\alpha, G) \\ \beta \quad & \sim \quad N(\mu, \textstyle\sum) \end{aligned} \quad (2)$$

where $PT(\boldsymbol{\alpha}, G)$ denotes a Pólya tree prior with prespecified parameters $\boldsymbol{\alpha}$ and G (see Section 3 for details) and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are prespecified parameters of the normal prior distribution on the coefficients $\boldsymbol{\beta}$.

A large number of suggestions have been made in the Bayesian survival analysis literature for modeling the baseline hazard (see Ibrahim et al., 2001, for descriptions). These schemes vary from parametric to nonparametric prior models, each introducing an additional level of complication in the statistical inference process, not to mention leaving a large number of candidate, and potentially complex, models from which the practitioner must choose. However, the baseline hazard is a nuisance parameter in our application or at the least in many such data analyses a parameter of secondary interest. Our suggestion is to thus "estimate it away" via an empirical Bayes type argument. By arguments similar to those in

Johansen (1983) and Klein (1992), the nonparametric estimator of $\Lambda_0(t)=\int_0^t \lambda_0(u)\,du$ is given by

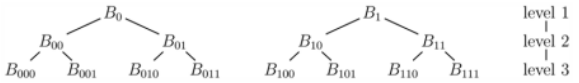
$$\widehat{\Lambda}_0(t|\widehat{\beta},\widehat{\mathbf{V}})=\int_0^t \frac{\sum_{i=1}^n \sum_{k=1}^{K_i} dN_{ik}(u)}{\sum_{i=1}^n \sum_{k=1}^{K_i} Y_{ik}(u)\exp\{\widehat{\beta}' Z_{ik}\}\widehat{V}_i} \quad (3)$$

where $N_{ik}(u) = I\{X_{ik} \le u, \delta_{ik} = 1\}$, $Y_{ik}(u) = I\{X_{ik} \ge u\}$, and $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{V}} = (\hat{V}_1, \cdots, \hat{V}_n)$ are, respectively, the expected values of $\boldsymbol{\beta}$ and $\mathbf{V}$ given the data and the current estimate of $\hat{\Lambda}_0$. As can be seen in Section 4, (3) can be readily modified to incorporate a Markov chain Monte Carlo sample $(\boldsymbol{\beta}_j, V_j)$, $j = 1, \cdots, J$.

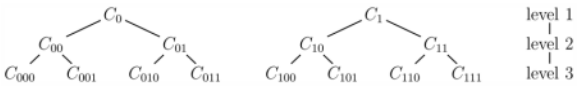
3 Pólya tree distribution

Walker and Mallick (1997) define the Pólya tree distribution as a prior for the frailty distribution, drawing heavily on the work of Lavine (1992, 1994); Ferguson (1974), Lavine (1992), and Mauldin, Sudderth, and Williams (1992) being the groundbreaking papers introducing the Pólya tree prior. We will not present the detailed mathematics in this section, but merely the intuition and notation for ease of exposition in the remainder of the paper, particularly the frailty distribution goodness of fit routines of Section 6.

The Pólya tree distribution constructs a distribution on the space of continuous probability distributions, with suitable choice of parameters, by cascading down a tree, each level of which partitions the domain of interest (e.g., the real line at finer and finer levels of detail). We partition the domain of interest into 2^m sets B_{ε_m} where ε_m is a binary sequence of length m.



We move through the tree via probabilities C_{ε_m}, corresponding analogously to B_{ε_m}. That is, each node of the C-tree below corresponds to the probability of entering the analogous node in the B-tree above.



For example, C_0 is the probability we start in B_0, $C_1 = 1-C_0$ is the probability we start in B_1. Upon entering B_{01}, with probability C_{010} we proceed to B_{010} and probability $C_{011} = 1 - C_{010}$ we proceed to B_{011}. In the ε-notation, upon entering B_{ε_m} at level m, with probability $C_{\varepsilon_m 0}$ we move to $B_{\varepsilon_m 0}$ and with probability $C_{\varepsilon_m 1}$ we move to $B_{\varepsilon_m 1}$.

Hanson (2006) introduces the notation $e_m(k) = \varepsilon_m$ for the kth set of the B-tree or C-tree at level m, $k = 1,\ldots, 2^m$. This notation provides simplification of the mathematical formulations later in the paper, as well as in coding the tree in the MCMC algorithms proposed, $e_m(k)$ being the binary representation of the number $k - 1$ by m digits. For example, $e_3(5) = 100$ in binary (i.e., the binary representation of the number four), identifies the fifth set in either the B-tree or C-tree at level 3.

Rather than fix the probabilities C_{ε_m}, we assume for each level m

$$(C_{\varepsilon_m 0}, C_{\varepsilon_m 1}) \sim beta(\alpha_{\varepsilon_m 0}, \alpha_{\varepsilon_m 1}) \quad (4)$$

where $\alpha_{\varepsilon_m 0}, \alpha_{\varepsilon_m 1} > 0$ are specified. We thus have a tree of α_{ε_m} values corresponding to each level of C_{ε_m}. Note though that $C_{\varepsilon_m 1} = 1 - C_{\varepsilon_m 0}$ for every sequence ε_m.

The bottom-line is that the probability of entering a node defined by the binary sequence ε_m at level m is the product of probabilities along the path traversing the tree to enter that node,

$$F(B_{\varepsilon_m}) = C_{\varepsilon_1} \cdots C_{\varepsilon_m}, \quad (5)$$

where ε_j contains the first j digits of ε_m, $j = 1,\ldots, m$. For example, if $\varepsilon_m = 100$ in binary, then $F(B_{100}) = C_1 \cdot C_{10} \cdot C_{100}$. Assuming the probabilities C_{ε_m} are independent as we progress down the levels of the C-tree, we may use (5) to simulate distributions F from the Pólya tree distribution. In order to implement such a simulation algorithm, presented below, we employ the following restrictions.

- To ease implementation, Lavine (1994) suggests sampling from a partially specified Pólya tree distribution, that is, a tree stopped at fixed level M. Walker and Mallick (1997) fix $M = 8$; Hanson and Johnson (2002) recommend $M \approx \log_2 n$ for sample size n.
- Following Walker and Mallick (1997) we assume at level M, each set $B_j = [G^{-1}\{(j-1)/2^M\}, G^{-1}\{j/2^M\}]$ for $j = 1,\ldots, 2^M$.
- Walker and Mallick (1997) overcomes problems in identifying an arbitrarily specified distribution over the space of frailty distributions by restricting the base measure G to have median zero. More specifically, they assume G is a normal distribution with mean zero and variance one hundred. Furthermore, $C_0 = C_1 = 0.5$.
- Following Lavine (1992) and Walker and Mallick (1997) we set $\alpha_{\varepsilon_m} = cm^2$ for $c > 0$.

In our simulations and applications, we will consider an appropriate specification of the variance of the base distribution G and the constant c in the specification of α_{ε_m}. As presented in Section 2, we will denote this Pólya tree distribution by $F \sim PT(\boldsymbol{\alpha}, G)$ where $\boldsymbol{\alpha}$ is the collection over all binary sequences ε_m, $m = 1,\ldots, M$, of α_{ε_m}.

Simulating a Pólya tree distribution random variate

1. Initialize G, M, $\{B_{\varepsilon_m}\}$, and $\{\alpha_{\varepsilon_m}\}$.
2. Set $C_0 = C_1 = 0.5$.
3. For each $m = 1,\ldots, M-1$,
   a. Generate $(C_{\varepsilon_m 0}, C_{\varepsilon_m 1}) \sim beta(\alpha_{\varepsilon_m 0}, \alpha_{\varepsilon_m 1})$,
   b. Compute $F(B_{\varepsilon_m})$ of equation (5) for each of the 2^M binary sequences ε_M.
4. For each $i = 1,\ldots, n$,
   a. Generate $U_i \sim Uniform(0, 1)$,
   b. Select interval k such that $\sum_{j=0}^{k-1} F(B_{\varepsilon_M(j)}) < U_i < \sum_{j=0}^{k} F(B_{\varepsilon_M(j)})$, $k = 1,\ldots, 2^M$,
   c. Generate $\theta_i \sim Uniform(B_{\varepsilon_M(k)})$.

Note that in step three, at level m we draw 2^{m-1} pairs of the probabilities C_{ε_m}. This algorithm in essence draws the tree in its entirety and computes the probability of entering any of the 2^M nodes in the last level M. In step 4b, we take $F(B_{\varepsilon_M(0)}) = 0$. In step 4c, if either of the first or 2^Mth extreme sets at level M is chosen, Walker and Mallick (1997) recommends sampling from the base measure restricted to that chosen extreme set, $G(B_{\varepsilon_M(1)})$ or $G(B_{\varepsilon_M(2_M)})$ respectively.

Pólya tree distributions admit a conjugacy result in that, conditional on the frailty parameters $\boldsymbol{\theta}$, for all ε_m,

$$(C_{\varepsilon_m 0}, C_{\varepsilon_m 1})|\theta \sim beta\,(\alpha_{\varepsilon_m 0}+n_{\varepsilon_m 0}(\theta),\ \alpha_{\varepsilon_m 1}+n_{\varepsilon_m 1}(\theta)) \quad (6)$$

where $n_{\varepsilon_m 0}(\boldsymbol{\theta})$ and $n_{\varepsilon_m 1}(\boldsymbol{\theta})$ are the number of frailties of $\boldsymbol{\theta}$ in the sets $B_{\varepsilon_m 0}$ and $B_{\varepsilon_m 1}$ respectively. Consequently, $F|\boldsymbol{\theta} \sim PT(\boldsymbol{\alpha}|\boldsymbol{\theta}, G)$, where $\boldsymbol{\alpha}|\boldsymbol{\theta}$ denotes the update of each α_{ε_m} through $n_{\varepsilon_m}(\boldsymbol{\theta})$.

4 Empirical Bayes Hastings sampler

The frailty model (1) with prior structure (2) does not lend to closed form posterior inference. We perform statistical inference via a Markov chain Monte Carlo method. We call this method the empirical Bayes Hastings sampler, in the spirit of Casella (2001), since each iteration of the sampler, we not only update the parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and F, but estimate the baseline cumulative hazard given current random variate generations and the data. In particular, at iteration r, the baseline hazard function is given by

$$\begin{aligned}\widehat{\Lambda}_0^{(r)}(t|\beta_1,\ldots,\beta_{J_r},\theta_1,\ldots,\theta_{J_r}) &= \arg\max_{\Lambda_0} \tfrac{1}{J_r}\sum_{j=1}^{J_r} l(\Lambda_0|\beta_j,\theta_j) \\ &= \tfrac{1}{J_r}\sum_{j=1}^{J_r}\int_0^t \frac{\sum_{i=1}^n\sum_{k=1}^{K_i} dN_{ik}(u)}{\psi_j(u)}\end{aligned} \quad (7)$$

where $\psi_j(u)=\sum_{i=1}^n\sum_{k=1}^{K_i} Y_{ik}(u)\exp\{\beta_j' Z_{ik}+\theta_{i;j}\}$ and $(\boldsymbol{\beta}_j, \boldsymbol{\theta}_j)$, $j = 1, \cdots, J_r$ are a sample from the Hastings sampler.

4.1 Algorithm

The Hastings sampler is as follows.

1. Initialize
   - $F^{(0)}$, $\boldsymbol{\beta}^{(0)}$, and $\boldsymbol{\theta}^{(0)}$;
   - $\widehat{\Lambda}_0^{(0)}$ via (3) using $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$;
   - base measure G.

   At iteration r
2. Set $\boldsymbol{\beta}_0 = \boldsymbol{\beta}^{(r-1)}$, $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^{(r-1)}$, $F_0 = F^{(r-1)}$.
3. Generate a sample $(\boldsymbol{\beta}_j, \boldsymbol{\theta}_j, F_j)$ for $j = 1,\ldots, J_r$, by iterating over the simulators

$$\beta_j \sim [\beta|\theta_{j-1}, F_{j-1}, \widehat{\Lambda}_0^{(r-1)}, data]$$
$$\theta_j \sim [\theta|\beta_j, F_{j-1}, \widehat{\Lambda}_0^{(r-1)}, data]$$
$$F_j \sim [F|\beta_j, \theta_j, \widehat{\Lambda}_0^{(r-1)}, data].$$

4. Compute the baseline hazard estimator $\widehat{\Lambda}_0^{(r)}$ from (7).
5. Set $\boldsymbol{\beta}^{(r)} = \boldsymbol{\beta}_{J_r}$, $\boldsymbol{\theta}^{(r)} = \boldsymbol{\theta}_{J_r}$, and $F^{(r)} = F_{J_r}$.
6. Repeat steps two through five until $\widehat{\Lambda}_0^{(r)}$ converge.
7. Repeat steps two and three to produce a final Hastings sample of size $J_r = M$.

Note that the full conditional distribution on $\boldsymbol{\theta}$ is the product of the full conditional distributions over each θ_i. We will thus update the univariate full conditional distributions on θ_i individually.

At each iteration r the algorithm fixes the baseline cumulative hazard at $\Lambda_0^{(r-1)}$ and generates a Hastings sample of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and F. Given these new variates, the algorithm obtains a new estimate of the baseline cumulative hazard, $\Lambda_0^{(r)}$, and then generates a new Hastings sample given this update. As noted by Casella (2001), step four is motivated by the Monte Carlo EM (MCEM) algorithm. We have two iterative processes: the Hastings sampler over j at each iteration r in step three and the MCEM algorithm over r iterating between the Hastings sampler and M-step estimate in steps three and four respectively. We have included steps two and five to delineate the two iterative properties from an algorithmic standpoint. In Section 4.4, we remark on a few implementation issues to consider in applying this algorithm.

4.2 Conditional distributions

Sampling from the conditional distributions in step three of the empirical Bayes Hastings sampler has been detailed in Walker and Mallick (1997) for the proposed Bayesian frailty model. In particular, $\boldsymbol{\beta}$ variates are generated via the polar slice sampler (see Appendix for details) and $\boldsymbol{\theta}$ variates are generated via a Metropolis-Hastings sampler with candidate samples from F. The conditional distribution on F is, in turn, a Pólya tree distribution with parameters updated according to the sampling of the frailties $\boldsymbol{\theta}$. These sampling schemes were found to work well in our application and do not rely on tuning parameters (e.g., proposal distribution parameters), lending easily to automation.

A final note on the role of the hazard function in the empirical Bayes Hastings sampler, recall that the likelihood is equal to, using the notation from Section 2,

$$\left\{\prod_{i,k}\lambda_0(t_{ik})^{\delta_{ik}}\right\}\exp\left(\sum_{i,k}\beta'\mathbf{Z}_{ik}\delta_{ik}\right)\exp\left(\sum_{i,k}\theta_i\delta_{ik}\right)\exp\left\{-\sum_{i,k}\Lambda_0(t_{ik})\exp(\beta'\mathbf{Z}_{ik}+\theta_i)\right\}. \quad (8)$$

The baseline hazard function $\lambda_0(t)$ is thus not required in any part of the Hastings sampler, canceling in the Hastings acceptance ratio for sampling $\boldsymbol{\theta}$ (and the acceptance ratio if a Hastings sampler is used to sample $\boldsymbol{\beta}$), a proportionality constant in the polar slice sampler for $\boldsymbol{\beta}$, and not appearing in the full conditional distribution on F. Thus the inconsistency of the baseline hazard rate (see Burr, 1994), is a non-issue as we require only the cumulative baseline hazard estimator, which is well behaved.

4.3 Convergence issues

The empirical Bayes Hastings sampler induces a Markov chain with stationary distribution being the posterior distribution of interest $\pi(\boldsymbol{\beta}, \boldsymbol{\theta}, F|\hat{\Lambda}_0, \mathbf{Y})$. The substitution of a frequentist estimate for the unknown cumulative baseline hazard each iteration of the sampler leads us slightly astray from the standard ergodic theory underlying MCMC samplers. The convergence proof for our empirical Bayes Hastings sampler follows the arguments of Casella (2001). However, the added complexity of estimating an unknown function (baseline hazard) in our hierarchical model, as compared to unknown hyperparameters, motivates us to briefly present the convergence theory of the empirical Bayes Hastings algorithm within the context of the frailty model.

First note that at iteration r, equation (7) suggests maximizing a Monte Carlo estimate of the conditional expectation $E\{l(\Lambda_0|\beta_1,\ldots,\beta_{J_r},\theta_1,\ldots,\theta_{J_r})|\Lambda_0^{(r-1)}\}$ using a sample $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{J_r}$, $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{J_r}$ from the distribution $\pi(\beta,\theta|\Lambda_0^{(r-1)},\boldsymbol{Y})$. This sample is obtained from the Hastings sampler in step three during the $(r-1)$st iteration. The empirical Bayes Hastings sampler may thus be thought of as an MCEM algorithm with an MCMC (Hastings sampler) E-step. We may then fall back on MCEM convergence theory (see for example Caffo, Jank, and Jones, 2005, and Fort and Moulines, 2003) to ensure estimates of the cumulative baseline hazard from step four in the algorithm converge to the maximum likelihood estimate $\hat{\Lambda}_0$.

Interest lies in posterior inferences on the parameters $\boldsymbol{\beta}$ and F. We must show that the stationary distribution of the Hastings sampler, estimating the unknown baseline hazard Λ_0 with $\hat{\Lambda}_0$, is the posterior distribution of interest $\pi(\boldsymbol{\beta}, \boldsymbol{\theta}, F|\Lambda_0, \mathbf{Y})$. The following theorem provides us with this result. Denote the true baseline hazard by Λ_0, the random parameters by $\boldsymbol{\Psi} = \{\boldsymbol{\beta}, \boldsymbol{\theta}, F\}$ for simplicity in notation, and the transition kernel for the empirical Bayes Hastings sampler using baseline hazard Λ_0 by $P(\cdot|\boldsymbol{\Psi}; \Lambda_0)$. This kernel thus denotes transitions from $\boldsymbol{\Psi}$ conditional on both the data $\mathbf{Y}$ and cumulative baseline hazard Λ_0; though note that we suppress indication of the data $\mathbf{Y}$ throughout. The convergence theory is studied through the total variation norm denoted $\|\cdot\|_{TV}$.

Theorem 4.1—Suppose that

- **a.** for fixed baseline hazard Λ_0, the Markov chain with transition kernel $P(\cdot|\boldsymbol{\Psi}; \Lambda_0)$ is ergodic with stationary distribution π;
- **b.** the transition kernel $P(\cdot|\boldsymbol{\Psi}; \Lambda_0)$ has the following property: for every $\varepsilon > 0$, there exists $\delta > 0$ such that for fixed baseline hazard functions $\Lambda_{0;1}$, $\Lambda_{0;2}$, over all times t, $|\Lambda_{0;1}(t) - \Lambda_{0;2}(t)| < \delta$ implies

$$\left\|\int P(\cdot|\boldsymbol{\Psi};\Lambda_{0;1})\,\mu_0(d\boldsymbol{\Psi}) - \int P(\cdot|\boldsymbol{\Psi};\Lambda_{0;2})\,\mu_0(d\boldsymbol{\Psi})\right\|_{TV} < \varepsilon$$

  for every initial distribution μ_0.

Then

$$\left\|\int P^{(r)}(\cdot|\boldsymbol{\Psi};\widehat{\Lambda}_0)\,\mu_0(d\boldsymbol{\Psi}) - \pi\right\|_{TV} \to 0 \; as \; r, \; n \to \infty$$

for every initial distribution μ_0.

Proof: Note to reviewers: we provide a detailed proof in the Appendix. We propose the appendix to appear as on online supplement to the paper.

By the triangle inequality we have that

$$\|\int P^{(r)}(\cdot|\boldsymbol{\Psi};\widehat{\Lambda}_0)\,\mu_0(d\boldsymbol{\Psi}) - \pi\|_{TV} \leq \|\int P^{(r)}(\cdot|\boldsymbol{\Psi};\widehat{\Lambda}_0)\,\mu_0(d\boldsymbol{\Psi}) - \int P^{(r)}(\cdot|\boldsymbol{\Psi};\Lambda_0)\,\mu_0(d\boldsymbol{\Psi})\|_{TV} + \|\int P^{(r)}(\cdot|\boldsymbol{\Psi};\Lambda_0)\,\mu_0(d\boldsymbol{\Psi}) - \pi\|_{TV}$$

for every initial distribution μ_0 and for fixed Λ_0. Condition (a) implies that the second term on the right hand side converges to zero. Condition (b), consistency of the estimator $\hat{\Lambda}_0$ from (3), and an induction argument (see Roberts, Rosenthal, and Schwartz, 1998) imply the first term on the right hand side converges to zero.

The asymptotic theory here is different than a standard MCMC convergence result in that we are studying convergence over the iteration count r, Monte Carlo sample size M, and sample size n. The MCEM argument ensures that steps three through six of the empirical Bayes Hastings sampler provides us with an MLE $\hat{\Lambda}_0$ for the baseline hazard asymptotically over r. Ergodic theory ensures that step seven of the algorithm provides us with a sample of size M from the posterior distribution $\pi(\boldsymbol{\Psi}|\hat{\Lambda}_0, \mathbf{Y})$. Theorem 4.1 states that given a consistent MCEM-type estimator of Λ_0, step seven induces a Markov chain with stationary distribution $\pi(\boldsymbol{\Psi}|\Lambda_0, \mathbf{Y})$ asymptotically over M and n.

For the algorithm put forth in Section 4.1 and our application in Section 8, the MCMC routine in that case is a Gibbs sampler over the parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\mathbf{F}$. The cumulative baseline hazard function enters the transition kernel through an exponential function presented in the likelihood (8). Consequently, condition (b) in Theorem 4.1 follows by a continuity argument.

4.4 Implementation issues

The Hastings sampler in Section 4.1 includes two iterative processes: a Hastings sampler in step three and an MCEM algorithm with a Markov chain Monte Carlo E-step in steps two through six. A few remarks are in order towards implementing each of these processes and the algorithm in general.

- Since step four is in essence a Monte Carlo EM M-step, we confront the problem of choosing an appropriate Monte Carlo sample size, J_r. In particular, Tanner (1993) suggests increasing the Monte Carlo sample size as the MCEM algorithm progresses since we require a more precise Monte Carlo E-step estimate as we enter smaller neighborhoods about the MLE. Thus the dependence of the Monte Carlo sample size J_r on the iteration count r. We adopt the method of Levine and Fan (2004) to choose the Monte Carlo sample size.
- The algorithm may be potentially computationally expensive requiring the implementation of a Hastings sampler each iteration r. We employ importance sampling techniques to overcome this expense, in particular, updating a single Hastings sample through importance weights at each iteration r, rather than generating a Hastings sample at each iteration (see Levine and Casella, 2001, for details).
- Steps two through six do not forgo the need for a burn-in in the Hastings sampler in step seven. In particular, it is well known that the Hastings sampler for fitting non-parametric Bayesian models with Pólya tree prior distributions mixes slowly

(Hanson, 2006). In the examples in the following section we choose a conservative burn-in and subsample the variates generated.

- We find specification of the spread of the base measure in the Pólya tree prior affects the resolution of the estimate of the frailty distribution of the model. We thus recommend gauging the range of the frailties through a fit of the Clayton and Cuzick (1985) extension of the Cox proportional hazards model with gamma frailty for multivariate survival times. Though we know this model is inappropriate for fitting our data, it is easily fit in software such as R/Splus and allows us to make conservative, ballpark guesses of the spread of the base measure.

5 Validation examples

In this section we perform simulation experiments to study the efficacy of our empirical and nonparametric Bayesian approaches for drawing inferences from the frailty model. We are particularly concerned with non-standard frailty distributions. We thus simulate data from two frailty models, apply the algorithm of Section 4.1, and study the parameter and frailty distribution estimates as a means of validating our modeling approach. The simulations consider frailty model (1) with

1. standard normal log frailty distribution $N(0, 1)$ with no covariates;
2. standard normal log frailty distribution $N(0, 1)$ and two covariates with $\boldsymbol{\beta} = (1, \ln 2)$;
3. mixture of normals log frailty distribution $0.5N(-1, 0.25) + 0.5N(1, 0.25)$ and no covariates;
4. mixture of normals log frailty distribution $0.5N(-1, 0.25) + 0.5N(1, 0.25)$ and two covariates with $\boldsymbol{\beta} = (1, \ln 2)$.

For both models 2 and 4, the first covariate takes a common value within each cluster: $z_{11} = z_{12} \sim$ Bernoulli(0.5), while the second covariate can take different values within the same cluster: (z_{21}, z_{22}) is distributed according to a bivariate normal distribution with zero vector mean, variance components of one, and correlation coefficient 0.75. The two covariates are independent of each other, and are also independent of the frailty distribution. We simulated data under each model with $n = 200$ clusters and $K_i = 2$ observations for each cluster (e.g., two eyes for each subject). Following an implementation of the Levine and Fan (2004) routine to allow the cumulative baseline hazard estimates to converge in steps two through six of Section 4.1, we ran the Gibbs sampler in step seven of the algorithm in Section 4.1 for 60,000 iterations, keeping every fifth sample, after a burn-in of 10,000 iterations.

Figure 1 presents the posterior estimates of the frailty distributions under each of these four simulation models. Note that the nonparametric modeling of the frailty distribution allows us to pick up the bell-shape and bimodal shapes of the distributions in each case, respectively.

Table 1 presents Monte Carlo estimates of the regression parameters $\boldsymbol{\beta}$ and standard deviations for each of the two simulation experiments in which covariates were included. These estimates are averages across 50 simulation runs for each model, that is, we are studying the estimator consistency. The algorithms are initiated with $\boldsymbol{\beta} = (0, 0)$. Note that in each case, the algorithm presents consistent estimates of the coefficients.

In order to study the gains in nonparametrically modeling the frailty distribution, we perform an additional simulation along the lines of frailty model number four above. However, we imagine that only one covariate is in the model, though correlated with another covariate (with correlation coefficient 0.5) that determines the mode of the frailty distribution. The algorithm specifications are the same as above. A normal log frailty fit to

the model finds a coefficient estimate of 0.39 with standard deviation 0.20. This estimate not only falls short of the true $\beta = 1$, but in fact infers a non-significant coefficient (i.e., not significantly different than zero at the 5% level). Figure 2 presents the posterior estimate of the frailty distribution in this simulation experiment. Note that the nonparametric modeling of the frailty distribution identifies two modes at −1 and 1. Furthermore, the inference on β is less affected by the unobserved covariate, presenting on average an estimate of 0.80 with standard deviation 0.11.

6 Goodness of fit for the frailty distribution

Testing for parametric forms of the frailty distribution within the nonparametric model (2) is relatively straightforward by nesting the hypothesized parametric frailty distribution within the Pólya tree prior. The goodness of fit may thus be evaluated through the Savage-Dickey ratio approximation of the Bayes factor (Verdinelli and Wasserman, 1995). Hanson (2006) summarizes computation of this approximate Bayes factor within Pólya tree models. In this section we will briefly detail the application of Hanson (2006) to the empirical Bayes Hastings sampler setting.

We may nest the hypothesized parametric frailty distribution within the Pólya tree prior distribution outlined in Section 3 by specifying the base measure G as this parametric frailty distribution. Lavine (1992) then shows that under the null hypothesis

$$H_0: C_{\varepsilon_j} = 0.5, \text{ for all } \varepsilon_j,\ j = 1, \ldots, M, \quad (9)$$

namely an equal probability of branching left or right down the Pólya tree, the frailty distribution F follows the parametric form of the base measure. Consequently, the goodness of fit test of this parametric frailty distribution is equivalent to testing hypothesis (9).

The Savage-Dickey density ratio approximates the Bayes factor of the posterior odds in favor of H_0 against the prior odds in favor of H_0 as

$$BF = \frac{P(H_0|Data)/P(H_A|Data)}{P(H_0)/P(H_A)} = \frac{\pi_{H_0}(\mathscr{C}|Data)}{\pi_{H_0}(\mathscr{C})}, \quad (10)$$

where $\pi_{H_0}(\mathscr{C})$ and $\pi_{H_0}(\mathscr{C}|Data)$ are the prior and posterior densities of $\mathscr{C}$, respectively, evaluated under the null hypothesis of all $C_{\varepsilon_j} = 0.5$. Here $\mathscr{C}$ represents the set of all elements in the C-tree, namely C_{ε_j} for all binary sequences ε_j, $j = 1, \ldots, M$, and $Data$ representing the triplet of failure/censoring time, censoring indicator, and covariates $\{X_{ik}, \delta_{ik}, Z_{ik}\}$ for each observation $k = 1, \ldots, K_i$ within cluster $i = 1, \ldots, n$. This formulation follows from nesting the hypothesized parametric form of the frailty distribution within the Pólya tree prior under which the prior distribution on the coefficients $\boldsymbol{\beta}$ in (2) is the same under the null and alternative hypotheses of interest (Verdinelli and Wasserman, 1995).

The prior density in (10) is easily computed under H_0 being the product of beta densities (4) evaluated at 0.5,

$$\pi_{H_0}(\mathscr{C}) = \prod_{k \in \{1,\ldots,2^j\};\ j \in \{0,\ldots,M-1\}} (0.5)^{\alpha_{e_{j(k)0}} + \alpha_{e_j(k)1} - 2} / Beta(\alpha_{e_j(k)0}, \alpha_{e_j(k)1})$$

where $Beta(\cdot, \cdot)$ is the Beta function and the first term in the product, at $j = 0$, uses parameters α_0 and α_1. Though we do not explicitly sample $\mathscr{C}$ in the Hastings sampler of Section 4, the posterior density may be computed from the variates of $\boldsymbol{\theta}$ generated

$$\pi_{H_0}(\mathscr{C}|Data) = (1/T)\sum_{i=1}^{T}\pi_{H_0}(\mathscr{C}|\beta^{(i)}, \theta^{(i)}, \Lambda_0, Data) = (1/T)\sum_{i=1}^{T}\pi_{H_0}(\mathscr{C}|\theta^{(i)}),$$

the Rao-Blackwellized estimator of Gelfand and Smith (1990) under the hierarchy in (1) and (2). Here

$$\pi_{H_0}(\mathscr{C}|\theta^{(i)}) = \prod_{k\in\{1,\ldots,2^j\};\, j\in\{0,\ldots,M-1\}} (0.5)^{\alpha_{e_{j(k)0}}(\theta^{(i)})+\alpha_{e_j(k)1}(\theta^{(i)})-2} / Beta(\alpha_{e_j(k)0}(\theta^{(i)}), \alpha_{e_j(k)1}(\theta^{(i)}))$$

where $\alpha_{e_j(k)0}(\boldsymbol{\theta}^{(i)}) = \alpha_{e_j(k)0} + n_{e_j(k)0}(\boldsymbol{\theta}^{(i)})$ and $\alpha_{e_j(k)1}(\boldsymbol{\theta}^{(i)}) = \alpha_{e_j(k)1} + n_{e_j(k)1}(\boldsymbol{\theta}^{(i)})$ with, as detailed in Section 3, the parameters $\alpha_{e_j(k)}(\boldsymbol{\theta}^{(i)})$ updated according to the variates $\boldsymbol{\theta}^{(i)}$ generated via $n_{e_j(k)}(\boldsymbol{\theta}^{(i)})$, the number of frailties of $\boldsymbol{\theta}^{(i)}$ in $B_{e_j(k)}$. Both the prior and posterior densities in the Bayes factor (10) involve merely a post-processing of the empirical Bayes Hastings sampler variates already generated to fit the frailty model, without any additional Hastings sampler iterations nor Hastings sampler implementations to generate variates from other marginal distributions.

Hanson (2006) notes that this Hastings sampler Savage-Dickey density ratio approximation to the Bayes factor is poor when the hypothesized frailty distribution is highly unlikely as the null value of $\mathscr{C}$ has very low posterior mass. However, this is a problem only when testing a hypothesized nonparametric frailty distribution. When testing a hypothesized parametric frailty distribution, a large Bayes factor is a large Bayes factor, the approximation not being thrown off to the point of incorrectly accepting a highly unlikely null hypothesis. Furthermore, the approximation performs well when the hypothesized parametric frailty distribution is moderately unfavored or favored by the data. On top of the minimal computational and coding expense in producing the Savage-Dickey density ratio (10), we thus propose this approach for goodness of fit testing of parametric frailty distributions within the empirical Bayes Hastings sampler framework.

7 Model selection

In this section we consider the problem of variable selection, choosing the optimal set of risk factors for predicting the outcome of interest. Comparisons between competing models may be made through Bayes factors. However, the Bayes factor computations of Section 6 are developed under the assumption of nested models, namely the hypothesized parametric form of the frailty distribution is nested within the nonparametric (Pólya tree prior) frailty distribution. In model selection, we may wish to compare models with non-intersecting sets of variables. Not only then are the models not nested, but the models are fitting nonparametric frailty distributions. Though we are not directly testing nonparametric alternatives to the frailty distribution, as mentioned at the end of Section 6, the Bayes factor approximation (10) may be unreliable. We thus take an alternative approach, using our Hastings sampler to simulate over the model space of interest.

Our goal in model selection is to decide whether each risk factor or covariate belongs in the final model, that is, whether the regression coefficient β_l for each $l = 1,\ldots, p$ in model (1) is zero or not. We follow the method of Gottardo and Raftery (2008), in which our Hastings sampler traverses over the model space by simulating zero or non-zero values for each regression coefficient (see also Kuo and Mallick, 1998). The coefficient value is chosen according to the Hastings acceptance rule, deciding whether to remove, say, variable l from the model ($\beta_l = 0$) or not ($\beta_l \neq 0$). To build such decisions into the Hastings sampler, we specify the prior distribution on each coefficient β_l as a mixture distribution

$$\beta_l \sim (1 - \pi_0) \cdot I(\beta_l=0)+\pi_0 \cdot \pi(\beta_l) \quad (11)$$

where π_0 is the prior probability variable l belongs in the model, $I(\beta_l = 0)$ is a point (Dirac) mass at zero, and $\pi(\beta_l)$ is the prior distribution on β_l from (2).

The prior distribution (11) is a mixture of mutually singular distributions, being the point mass at zero and the continuous distribution $\pi(\beta_l)$. However, Gottardo and Raftery (2008) show that such mixture priors present minimal difficulties as we may obtain a density with respect to the measure $(\delta_0 + \lambda)$, the sum of the Dirac mass at zero and the Lebesgue measure. In our case, the prior density is

$$(1 - \pi_0) \cdot I(\beta_l=0)+\pi_0 \cdot \varphi(\beta_l|0, \sigma^2_{\beta_l}) \cdot I(\beta_l \neq 0)$$

where $\varphi(\beta_l|0, \sigma^2_{\beta_l})$ is the normal density on β_l with mean zero and variance $\sigma^2_{\beta_l}$ as derived from the multivariate normal prior in (2). Note that this mixture density explicitly models the inclusion and exclusion of variable l from the model, through components corresponding to $\beta_l = 0$ and $\beta_l \neq 0$. The removal of zero from the support of the second component, nonetheless, is required to ensure a valid density with respect to the measure $(\delta_0 + \lambda)$.

The polar slice sampler (see Section 4.2 and the Appendix) is easily modified to handle the mixture prior. However, we recommend performing the model selection with the mixture prior distribution (11) for a “burn-in” period. Upon achieving equilibrium, the best model is chosen and the sampler presented in Section 4 for prior distribution (2) is run to draw inferences and study frailty distribution goodness-of-fit.

Gottardo and Raftery (2008) discuss a Metropolis-Hastings sampler implementation of MCMC over mixtures of mutually singular distributions which may be adapted within our setting. Gottardo and Raftery (2008) find that such an implementation is computationally more expensive, statistically less efficient, and less easily automated as compared to the Gibbs sampler (polar slice sampler) implementation. We thus choose the Gibbs steps as they fit naturally within our proposed empirical Bayes Hastings sampler and performs admirably in the SWAP application. However, in applications where the requisite full conditional distribution are not available for the polar slice sampler, the Metropolis-Hastings sampler provides a viable alternative.

8 Analysis of the SWAP data

In this section we illustrate the proposed empirical Bayes Hastings sampler for frailty modeling on a data set collected by Demirel and Johnson (2001) to study the performance of two methods for detecting early glaucomatous visual field damage, standard automated perimetry (SAP) and short wavelength automated perimetry (SWAP). The data consists of 220 subjects (440 eyes) with ocular hypertension, as measured by an untreated intraocular

pressure (IOP) greater than 21 mm Hg, at two occasions during the baseline period, SAP fields within normal limits at baseline relative to an age representative pool of normal individuals, and no other condition that may affect visual fields other than the risk of glaucoma, recruited from the Sacramento, California region. The data set under study here is slightly smaller than that used in Demirel and Johnson (2001) as we included only subjects with vertical cup-to-disc data. As such data is an important, and in this study the only, measure of structure towards glaucomatous end point, we deemed it a pre-requisite for our analysis. The subjects were followed for at least three years, with median follow up of four years and maximum follow up 9.6 years. The average baseline IOP across the subjects was 23.87 mm Hg with standard deviation 2.94 mm Hg. The outcome measure is glaucomatous visual field loss determined by a classification of abnormal visual fields based on SAP on two consecutive visits during the study period. Eighteen eyes (4% of study eyes) reached endpoint during the follow up period. These eighteen eyes come from fourteen subjects (6% of study subjects), namely four subjects had both eyes reach end point.

Demirel and Johnson (2001) consider the prevalence and incidence of SWAP and SAP deficits over the period of the study. However, for the purposes of this illustration, we consider the relationship of baseline SWAP results (BLSWAP) as well as SAP-based visual field, clinical, and demographic measures on glaucomatous visual field abnormality based on SAP. Table 2 summarizes all baseline predictor variables in the analysis. Note that all ocular variables, except family history of glaucoma, are measured for each eye (eye-specific), whereas demographic variables are patient-specific. The subjects in the study had an average age of 57, ranging from a 13 year old to an 81 year old; 113 subjects were male; 20 subjects were African-American; and 58 subjects identified a family history of glaucoma.

The times to glaucomatous visual field loss for the two eyes from each patient can be expected to be correlated with unknown dependence structure influenced by unobserved co-variates (e.g., treatment strategies for ocular hypertensive patients and more detailed visual field and optic disc measurements/imaging). Specifying a parametric frailty distribution, for example the popular gamma distribution, thus seems arbitrary at best, motivating the use of a nonparametric frailty term.

We use the routine of Section 7 to select an appropriate model for the relationship between the risk factors of Table 2 and onset of glaucoma. In this model selection routine, we force the baseline SWAP indicator (within or outside normal limits) into the model as the SWAP indicator is the variable of key interest in the study. SWAP is believed to be able to detect early glaucomatous abnormalities several years before SAP (Demirel and Johnson, 2001). Hence the baseline SWAP status may be predictive of the confirmed glaucomatous endpoint based on SAP. We also force VCD and IOP into the model since VCD is the only optic disc measurement collected in the study and since study subjects all have ocular hypertension. Thus VCD and IOP are important covariates over which to control in a model of time to glaucomatous visual field loss. Variable selection is thus performed over all other variables in the data set. We set the parameters in the prior distribution on $\boldsymbol{\beta}$ to be the estimates from a fitted frequentist gamma frailty model to the relevant set of covariates. We initialize $\boldsymbol{\beta}^{(0)}$ at the prior mean, $\boldsymbol{\theta}^{(0)}$ with a zero vector, $F^{(0)}$ at a sample from the Pólya tree distribution with parameters α_{e_m}, and $\widehat{\Lambda}_0^{(0)}$ at the estimate (3) with the initial values of $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$. The prior probability that a variable is included in the final model is taken as $\pi_0 = 0.5$, so every model is equally likely a priori.

The last column of Table 2 presents the posterior probability that each regression co-efficient, β_l, $l = 1,\ldots, 11$, is non-zero. Table 3 presents the posterior probability of each model selected. Note that the best model selected includes the variables MD, gender, age, BLSWAP, IOP, and VCD. Subject gender and age, though not forced into the model, appear

in all models presented in Table 3. Interestingly, mean deviation is the variable selected to account for visual field damage on glaucomatous loss, as opposed to either PSD or CPSD. We note that the additional steps of simulating each regression coefficient separately as part of the model selection routine within the Hastings algorithm requires negligible computational cost and minimal additional coding. As an aside, though perhaps obvious, we found coding the models as 11-digit binary numbers, with digit l identifying whether variable $l = 1,\ldots, 11$ is in the model (one) or not in the model (zero), facilitates summaries of model posterior probabilities through conversions back and forth between the binary and decimal representations. For example, the best model is coded as 10001011011 with decimal representation of 1,115.

To test the hypothesis of a gamma frailty distribution, we assume a log-gamma base measure G on the log-frailty parameter (θ_i) for the Pólya tree distribution prior. As suggested by Walker and Mallick (1997) and Lavine (1992), we center the first partition of the Pólya tree at the zero. This assumption is analogous to fixing the gamma frailty distribution to a mean of one, the motivation here being to fix the median of the frailty distribution F to ensure identifiability of the frailty distribution from the baseline hazard in our model. A fit of a frequentist gamma frailty model to the data suggests a standard deviation of 8.4 for the distribution of the frailties. We grow the tree to $M = \log_2 220 = 8$ levels assuming $\alpha_{\varepsilon_m} = c \cdot m^2$ with $c = 1$ as recommended by Hanson and Johnson (2002). We note as well that inferences drawn in this application appear to be robust to choice of the partition tree, both specification of the base measure and options for α_{ε_m} that are less susceptible to over-smoothing. We refer the reader to the Appendix Section 10.3 for results from a sensitivity analysis to this end.

Table 4 presents descriptive summaries of the final model selected. Figure 3 presents the estimated frailty distribution. The routine of Section 6 obtains a Bayes factor on the order of 10^{12} against a hypothesized gamma frailty distribution. The bimodal nature of the frailty distribution suggests the possible absence of significant covariates in the prediction of glaucoma onset. The SWAP indicator turns out not to be a significant predictor of glaucomatous endpoint at the 5% “significance level.” Though moderately surprising, this finding may be explained by three factors. First, as mentioned earlier, SWAP has been shown to detect glaucomatous abnormalities on the order of 3 to 5 years earlier than SAP. However, our data considers a 10 year follow-up and time to endpoint of up to 8.2 years. Baseline SWAP can not be expected to predict an endpoint that far in advance, especially when there are other strong predictors in the model. Second, visual field tests are highly variable, glaucomatous endpoints typically determined by at least two consecutive abnormal fields. Thus a single baseline SWAP evaluation as used in this model may not be sufficient. Our analysis thus suggests expanding the study and modeling machinery to longitudinally collected SWAP and abnormalities detected by consecutive, multiple SWAP evaluations (office visits). Third, we note that this data set contains only 18 endpoints from which we aimed to draw inferences on glaucomatous progression and the predictive ability of SWAP. Nonetheless, though the power, per se, of the inferential procedures, is low, the “significant” findings suggest a signal which may be reconfirmed in future analyses with more outcomes.

9 Discussion

The Bayesian frailty model approach put forth is modular with respect to nonparametric, parametric, or empirical Bayes considerations for the three components: regression parameters, frailty distribution, and cumulative baseline hazard. For example, empirical Bayes methods might be considered for a “nuisance” frailty (parametric) distribution. And if the cumulative baseline hazard or baseline hazard rate is of interest, a prior on and MCMC updating of that function may be considered. Of course, care must be taken in choice of such

a frailty distribution as incorrect specification may affect the regression coefficient estimates. Furthermore, as suggested by a reviewer, an alternative Pólya tree model approach such as finite mixtures of Pólya trees (see e.g., Hanson, 2006), random Pólya tree (Paddock et al., 2003), or optional Pólya trees (Wong and Ma, 2010) may be considered for the distribution of the log-frailty parameters. As the sensitivity analysis in the Appendix Section 10.3 suggest a robustness to choice of tree partitioning in our application, one of the primary reasons to opt for these alternatives, we leave extension of our empirical Bayes Hastings sampler scheme to that end for future work.

Nonetheless, the empirical Bayesian Hastings sampler proposed outputs a frequentist estimate of the cumulative baseline hazard function, $\hat{\Lambda}_0(t)$. Though we motivate the approach thinking of this function as a nuisance parameter in our application or of secondary interest in many such analysis problems, survival and hazard curves over desired covariate combinations may be presented along with an estimate of variability.

One of the primary motivations for developing our empirical Bayes Hastings method is the potential for automation of the routine, in comparison say to the popular piecewise-constant hazard (gamma process) approach used by Clayton (1991) and Walker and Mallick (1997), being computationally no more expensive and easier to code through a computation of the frequentist nonparametric cumulative hazard estimator instead of a Gibbs sampler. Furthermore, we are alleviated from any tuning parameter tweaking, for example, in the case of the piecewise-constant hazard selection of the prior process parameters and the number of independent increments or “pieces”. Granted Walker and Mallick (1997) argues inferences drawn are insensitive to sensible choices of these parameters, Clayton (1991) stating that the piecewise-constant prior model for Λ_0 is adequate when the regression parameters are of primary interest. However, the independent-increments prior process, as the name suggests, creates highly discretized and independent hazards. Such disadvantages may be overcome, but even a simple extension to the correlated prior process of Aslanidou, Dey, and Sinha (1998) requires application of the Metropolis-Hastings algorithm significantly increasing computational expense in the form of coding, tweaking, and run time. Of course, further generalizations as put forth in Chapters 3 and 4 of Ibrahim et al. (2001) merely exacerbate the situation comprising computational expense for realism, a trade off we argue is unnecessary with the empirical Bayes Hastings sampler approach.

Acknowledgments

We thank Dr. Chris Johnson for sharing the SWAP data analyzed in the paper.

References

1. Aslanidou H, Dey DK, Sinha D. Bayesian Analysis of Multivariate Survival Data Using Monte Carlo Methods. The Canadian Journal of Statistics. 1998; 26:33–48.
2. Burr D. On Inconsistency of Breslow’s Estimator as an Estimator of the Hazard Rate in the Cox Model. Biometrics. 1994; 50:1142–1145. [PubMed: 7786994]
3. Caffo BS, Jank W, Jones GL. Ascent-Based Monte Carlo EM. Journal of the Royal Statistical Society, Series B. 2005; 67:235–252.
4. Casella G. Empirical Bayes Gibbs Sampling. Biostatistics. 2001; 2:485–500. [PubMed: 12933638]
5. Clayton DG. A Monte Carlo Method for Bayesian Inference in Frailty Models. Biometrics. 1991; 47:467–485. [PubMed: 1912256]
6. Clayton DG, Cuzick J. Multivariate generalization of the proportional hazards model. Journal of the Royal Statistical Society, Series A. 1985; 148:82–108.
7. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society B. 1972; 34:187–202.

8. Demirel S, Johnson CA. Incidence and prevalence of short wavelength automate perimetry deficits in ocular hypertensive patients. American Journal of Ophthalmology. 2001; 131 (6):709–715. [PubMed: 11384565]
9. Ferguson TS. Prior distributions on spaces of probability measures. Annals of Statistics. 1974; 2:615–629.
10. Fort G, Moulines E. Convergence of the Monte-Carlo EM for curved exponential families. Annals of Statistics. 2003; 31:1220–1259.
11. Gelfand AE, Smith AFM. Sampling Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association. 1990; 85:398–409.
12. Gottardo R, Raftery AE. Markov Chain Monte Carlo with Mixtures of Mutually Singular Distributions. Journal of Computational and Graphical Statistics. 2009; 17:949–975.
13. Gustafson P, Aeschliman D, Levy AR. A Simple Approach to Fitting Bayesian Survival Models. Lifetime Data Analysis. 2003; 9:5–19. [PubMed: 12602771]
14. Hanson TE. Inference for Mixtures of Finite Polya Tree Models. Journal of the American Statistical Association. 2006; 101:1548–1565.
15. Hanson TE, Johnson WO. Modeling Regression Error with a Mixture of Polya Trees. Journal of the American Statistical Association. 2002; 97:1020–1033.
16. Ibrahim, JG.; Chen, M-H.; Sinha, D. Bayesian Survival Analysis. Springer; New York: 2001.
17. Johansen S. An Extension of Cox's Regression Model. International Statistical Review. 1983; 51:165–174.
18. Klein JP. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. Biometrics. 1992; 48:795–806. [PubMed: 1420842]
19. Kuo L, Mallick B. Variable Selection for Regression Models. Sankyâ. 1998; 60:65–81.
20. Lavine M. Some Aspects of Polya Tree Distributions for Statistical Modelling. Annals of Statistics. 1992; 20:1222–1235.
21. Lavine M. More Aspects of Polya tree Distributions for Statistical Modelling. Annals of Statistics. 1994; 22:1161–1176.
22. Levine RA, Casella G. Implementations of the Monte Carlo EM Algorithm. Journal of Computational and Graphical Statistics. 2001; 10:422–439.
23. Levine RA, Fan J. An Automated Monte Carlo EM Algorithm. Journal of Statistical Computation and Simulation. 2004; 74:349–359.
24. Mauldin RD, Sudderth WD, Williams SC. Polya trees and random distributions. Annals of Statistics. 1992; 20:1203–1221.
25. Paddock SM, Ruggeri F, Lavine M, West M. Randomized Polya tree models for nonparametric Bayesian inference. Statistica Sinica. 2003; 13:443460.
26. Robert, CP.; Casella, G. Monte Carlo Statistical Methods. 2. Springer; New York: 2004.
27. Roberts GO, Rosenthal JS. The polar slice sampler. Stochastic Models. 2002; 18:257–280.
28. Roberts GO, Rosenthal JS, Schwartz PO. Convergence Properties of Perturbed Markov Chains. Journal of Applied Probability. 1998; 35:1–11.
29. Tanner, MA. Tools for Statistical Inference. 3. Springer; NY: 1993.
30. Walker SG, Mallick BK. Hierarchical Generalized Linear Models and Frailty Models with Bayesian Nonparametric Mixing. Journal of the Royal Statistical Society. 1997; 59:845–860.
31. Wong WH, Ma L. Optional Pólya tree and Bayesian inference. Annals of Statistics. 2010; 38:14331459.
32. Verdinelli I, Wasserman L. Computing Bayes Factors using a Generalization of the Savage-Dickey Density Ratio. Journal of the American Statistical Association. 1995; 90:614–618.

10 Appendix

We propose this appendix to appear in an online supplement for the paper.

10.1 Sampling regression parameters

To sample from the full conditional distribution [$\boldsymbol{\beta}|\boldsymbol{\theta}$, F, $\hat{\Lambda}_0$, *data*], we apply the polar slice sampler of Roberts and Rosenthal (2002), see also Robert and Casella (2004, Chapter 8). This sampler augments the target variates $\boldsymbol{\beta} = (\beta_1,\ldots, \beta_p)$ to produce a Gibbs sampler over these p variates and the augmented variates. However the conditional distributions are available in closed form, the only complication being a restricted support, easily lending to an automated sampling routine which, for our application, works without any difficulties.

Consider sampling from one component of $\boldsymbol{\beta}$, say β_1 for illustration. In the remainder, we suppress the conditioning arguments $\{\boldsymbol{\theta}, F, \hat{\Lambda}_0, \boldsymbol{\beta}_{-1}, \textit{data}\}$, where $\boldsymbol{\beta}_{-1}$ is $\boldsymbol{\beta}$ without the first component. Introduce variables U and V on the positive real line such that

$$f(\beta_1, u, v) \propto I\{u<\mathscr{A}(\beta_1), v<\mathscr{B}(\beta_1)\} \cdot \pi(\beta_1|\beta_2, \ldots, \beta_p)$$

where

$$\mathscr{A}(\beta_1)=\exp\left(\sum_{ik}\delta_{ik}\beta_1 X_{ik1}\right),$$
$$\mathscr{B}(\beta_1)=\exp\left\{\sum_{ik}\widehat{\Lambda}_0(t_{ik})\exp(\theta_i)\right\}\exp\left(\sum_{l=2}^{p}\beta_l X_{ikl}\right)\exp(\beta_1 X_{ik1}),$$

X_{ikl} is the lth element of the vector $\mathbf{X}_{ik}$, $\pi(\beta_1|\beta_2,\ldots, \beta_p)$ is the full conditional distribution of β_1 from the prior distribution on $\boldsymbol{\beta}$ (in our case a univariate normal distribution), and the constrained space over which u and v are defined derive from the likelihood (8). We may then implement a Gibbs sampler iterating over the full conditional distributions

$$u|v, \beta_1 \sim \textit{uniform}\{0,\ \mathscr{A}(\beta_1)\}$$
$$v|u, \beta_1 \sim \textit{uniform}\{0,\ \mathscr{B}(\beta_1)\}$$
$$\beta_1|u, v \sim \textit{normal}\{\mu_1+\mathbf{v}_{12}\mathbf{V}_{22}^{-1}(\beta_{-1}-\mu_{-1}), w_{11}^{-1}\}, \text{with}\,\beta_1 \text{ satisfying } u<\mathscr{A}(\beta_1),\ v<\mathscr{B}(\beta_1)$$

where $\boldsymbol{\mu}_{-1}$ is the $\boldsymbol{\beta}$ prior mean vector without the first component,

$$\sum=\begin{pmatrix} v_{11} & \mathbf{v}_{12} \\ \mathbf{v}_{12}' & \mathbf{V}_{22}\end{pmatrix},\ \text{and}\ \mathbf{W}=\sum^{-1}=\begin{pmatrix} w_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{12}' & \mathbf{W}_{22}\end{pmatrix}.$$

Sampling from $\boldsymbol{\beta}$ thus requires sampling from $3p$ conditional distributions, the full conditional distributions on each of the p components of $\boldsymbol{\beta}$ augmented by draws of two variates, u and v, for each component.

For the polar slice sampler required in Section 7, the prior distribution is specified componentwise lending direct application of the above Gibbs sampler though generating variates, say for the first component as illustrated, from the prior mixture distribution on β_1.

10.2 Proof of Theorem 4.1

We will need the following lemma.

Lemma 10.1

If condition (b) of Theorem 4.1 holds, then the r-step transition kernel $P^{(r)}(\cdot|\mathbf{\Psi};\Lambda_0)$ *has the same property: for every* $\varepsilon > 0$, *there exists* $\delta > 0$ *such that for fixed baseline hazard functions* $\Lambda_{0;1}$, $\Lambda_{0;2}$, *over all times t,* $|\Lambda_{0;1}(t) - \Lambda_{0;2}(t)| < \delta$ *implies*

$$\|\int P^{(r)}(\cdot|\mathbf{\Psi};\Lambda_{0;1})\,\mu_0(d\mathbf{\Psi}) - \int P^{(r)}(\cdot|\mathbf{\Psi};\Lambda_{0;2})\,\mu_0(d\mathbf{\Psi})\|_{TV} < \varepsilon$$

for every initial distribution μ_0.

Proof—The proof follows the induction argument of Roberts et al. (1998). Let $\varepsilon > 0$ and pick $\delta > 0$ such that for all times $t\,|\Lambda_{0;1}(t) - \Lambda_{0;2}(t)| < \delta$. Note that for fixed Λ_0^*, $P(\cdot|\mathbf{\Psi};\Lambda_0^*)$ is a transition kernel of a Hastings sampler with stationary distribution π. For $r \geq 2$ and for any set A, we also have

$$\begin{aligned} P^{(r)}(A|\mathbf{\Psi};\Lambda_{0;1}) - P^{(r)}(A|\mathbf{\Psi};\Lambda_{0;2}) &= \int P^{(r-1)}(A|\eta;\Lambda_{0;1})P(\eta|\mathbf{\Psi};\Lambda_{0;1})\,d\eta - \int P^{(r-1)}(A|\eta;\Lambda_{0;2})P(\eta|\mathbf{\Psi};\Lambda_{0;2})\,d\eta \pm \int P^{(r-1)}(A|\eta;\Lambda_{0;1})P(\eta|\mathbf{\Psi};\Lambda_{0;2})\,d\eta \\ &= \int P^{(r-1)}(A|\eta;\Lambda_{0;1})\,\{P(\eta|\mathbf{\Psi};\Lambda_{0;1}) - P(\eta|\mathbf{\Psi};\Lambda_{0;2})\}\,d\eta + \int \{P^{(r-1)}(A|\eta;\Lambda_{0;1}) - P^{(r-1)}(A|\eta;\Lambda_{0;2})\}P(\eta|\mathbf{\Psi};\Lambda_{0;2})\,d\eta \\ &\leq \left\|P(\cdot|\mathbf{\Psi};\Lambda_{0;1}) - P(\cdot|\mathbf{\Psi};\Lambda_{0;2})\right\|_{TV} + \int \left\|P^{(r-1)}(\cdot|\mathbf{\Psi};\Lambda_{0;1}) - P(r-1)(\cdot|\mathbf{\Psi};\Lambda_{0;2})\right\|_{TV} P(\eta|\mathbf{\Psi};\Lambda_{0;2})\,d\eta \end{aligned}$$

uniformly in A. Thus by condition (b) and the inductive hypothesis, we have that

$$\left\|P^{(r)}(\cdot|\mathbf{\Psi};\Lambda_{0;1}) - P^{(r)}(\cdot|\mathbf{\Psi};\Lambda_{0;2})\right\|_{TV} < \varepsilon.$$

We may now prove Theorem 4.1. By the triangle inequality we have that

$$\|\int P^{(r)}(\cdot|\mathbf{\Psi};\widehat{\Lambda}_0)\,\mu_0(d\mathbf{\Psi}) - \pi\|_{TV} \leq \|\int P^{(r)}(\cdot|\mathbf{\Psi};\widehat{\Lambda}_0)\,\mu_0(d\mathbf{\Psi}) - \int P^{(r)}(\cdot|\mathbf{\Psi};\Lambda_0)\,\mu_0(d\mathbf{\Psi})\|_{TV} + \|\int P^{(r)}(\cdot|\mathbf{\Psi};\Lambda_0)\,\mu_0(d\mathbf{\Psi}) - \pi\|_{TV}$$

for every initial distribution μ_0 and for fixed Λ_0.

By consistency of the estimator $\hat{\Lambda}_0$ from (3), there exists an R such that for all $r \geq R$,

$$\|\int P^{(r)}(\cdot|\mathbf{\Psi};\Lambda_0)\,\mu_0(d\mathbf{\Psi}) - \pi\|_{TV} < \varepsilon/2.$$

Let γ, $\delta > 0$. By Egoroff's Theorem and consistency of $\hat{\Lambda}_0$ from (3), there exists an N such that for all times t, $|\hat{\Lambda}(t) - \Lambda_0(t)| < \delta$ for all $n \geq N$ except on a set of π-measure less than γ. Therefore, by condition (b), for all $\varepsilon > 0$, there exists an R_n for each n such that for times t, $|\Lambda_0^*(t) - \Lambda_0(t)| < \delta$ implies

$$\left\| \int P^{(r)}(\cdot|\boldsymbol{\Psi};\widehat{\Lambda}_0^*)\,\mu_0(d\boldsymbol{\Psi}) - \int P^{(r)}(\cdot|\boldsymbol{\Psi};\Lambda_0)\,\mu_0(d\boldsymbol{\Psi}) \right\|_{TV} < \varepsilon/2$$

for all $r \geq R_n$. Consequently, for every initial distribution μ_0 and all $r \geq \max\{R, R_n\}$,

$$\left\| \int P^{(r)}(\cdot|\boldsymbol{\Psi};\widehat{\Lambda}_0)\,\mu_0(d\boldsymbol{\Psi}) - \int P^{(r)}(\cdot|\boldsymbol{\Psi};\Lambda_0)\,\mu_0(d\boldsymbol{\Psi}) \right\|_{TV} \leq \sup_{\Lambda_0^*:\left\|\Lambda_0^*-\Lambda_0\right\|<\delta} \left\| \int P^{(r)}(\cdot|\boldsymbol{\Psi};\Lambda_0^*)\,\mu_0(d\boldsymbol{\Psi}) - \int P^{(r)}(\cdot|\boldsymbol{\Psi};\Lambda_0)\,\mu_0(d\boldsymbol{\Psi}) \right\|_{TV} < \varepsilon/2.$$

10.3 Sensitivity analysis: tree partitioning

The Pólya tree prior used in this paper is known to potentially leave inferences highly dependent on the tree partition chosen. In this section, we present a subset of results from a sensitivity analysis to study the impact of construction of the B- and C-trees of Section 3, specifically specification of the sets $B_{\varepsilon m}$ through the chosen base measure and the beta distribution in (4) through choice of the $\alpha_{\varepsilon m}$ parameters. For the latter, a reviewer correctly pointed out that the specification $\alpha_{\varepsilon m} = cm^2$ at level $m = 1,\ldots, M$ with $c > 0$ may over-smooth and hide dependence on the partition for higher (low m) levels of the tree. We thus choose a slower decay for small m setting $\alpha_{\varepsilon m} = c \cdot m\mathbb{I}(m \leq 4) + c \cdot m^2\mathbb{I}(m \geq 5)$ in the following analysis.

For the B-tree, we vary the partition here by fitting the model with two base measures: gamma distribution and normal distribution, varying the standard deviation of each for a "tight" partition, "moderate" partition analogous to that used in the analysis of Section 8, and a "wide" partition equating to small, medium, and large standard deviations of 1, 8, and 15 respectively. In each case, we ran our empirical Bayes Hastings sampler on the SWAP in an analogous manner to that performed in Section 8, the only modification being in the specification of the $\alpha_{\varepsilon m}$, as noted in the previous paragraph, and the base measure chosen. Figure 4 presents descriptive summaries of the posterior distribution on $\boldsymbol{\beta}$ from the six base measure scenarios considered. The figure suggests inferences are robust to choice of tree partition, the posterior means and standard deviations varying little for each of the six variables gender, age, intraocular pressure, vertical cup-to-disc ratio, mean deviation, and SWAP indicator. Furthermore, not only are the 95% credible sets for each coefficient of similar range across different base measure and standard deviation choices, but, correspondingly, do not suggest any coefficient estimates as outlying nor changes in "statistical significance". We note that a vast array of options have been considered in addition to the scenarios presented here, inferences being robust over all analyses performed.

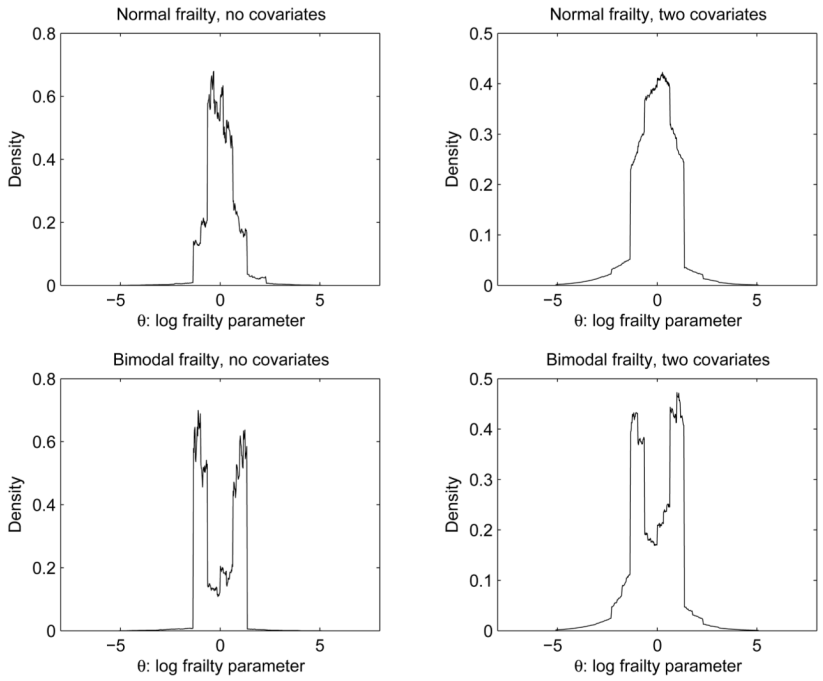


Figure 1.
Empirical Bayes Hastings sampler estimated frailty distributions from each of the four simulation experiments.

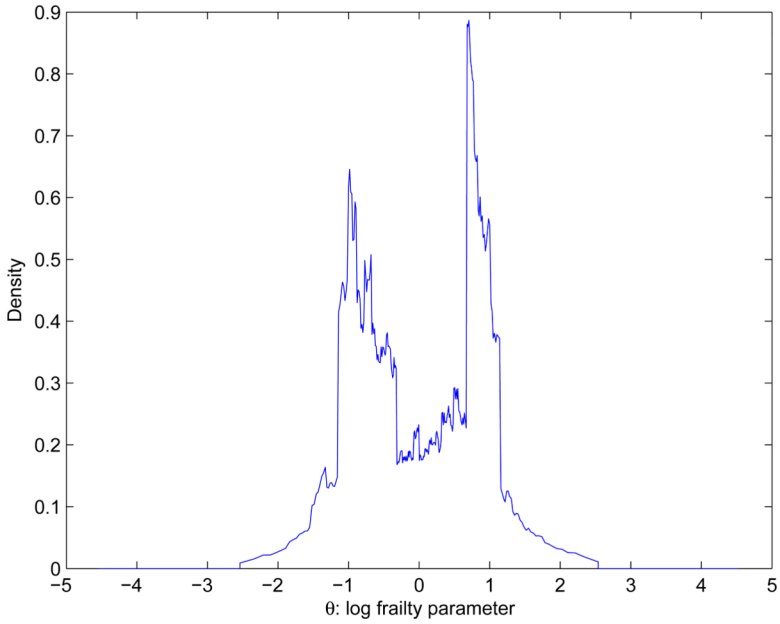


Figure 2.
Empirical Bayes Hastings sampler estimated frailty distribution from the simulation experiment with frailty distribution being a mixture of normals and unobserved covariate.

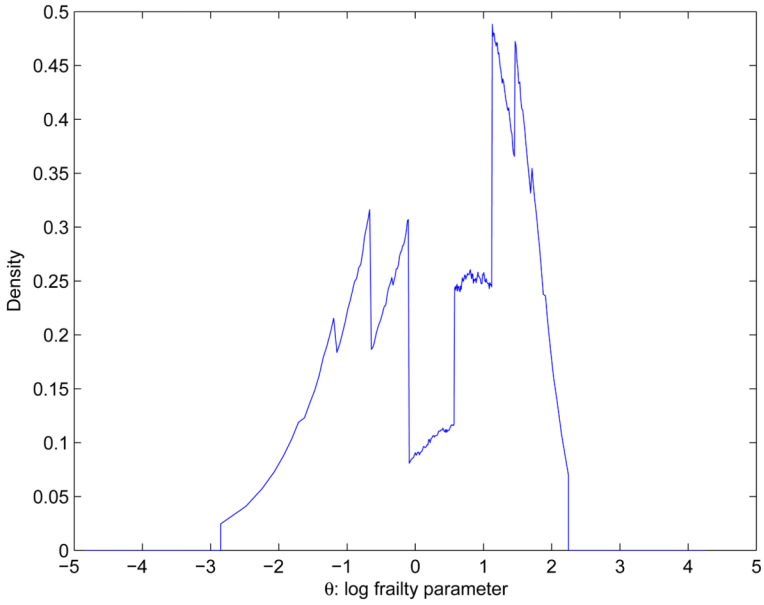


Figure 3.
Empirical Bayes Hastings sampler estimated frailty distribution over the six regression coefficients corresponding to the model presented in Table 4 for the SWAP data.

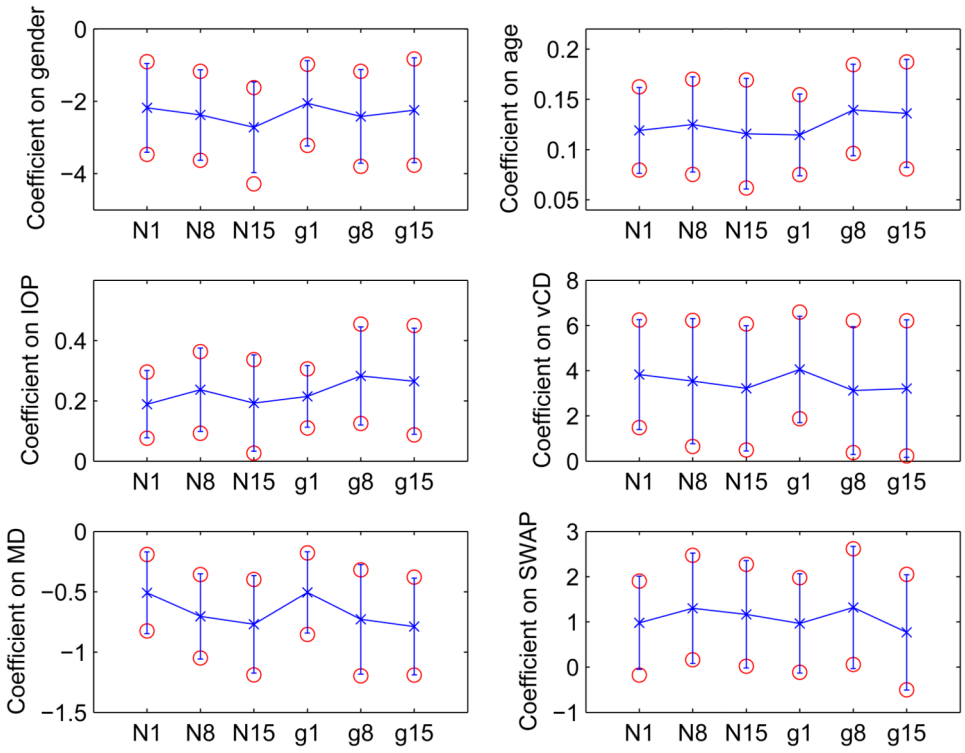


Figure 4. This figure will appear as part of the online supplement to accompany appendix Section 10.3

Summaries of the posterior distribution of $\boldsymbol{\beta}$ in study of six scenarios for choice of base measure (normally distributed base measure with standard deviations one, eight, and fifteen, N1, N8, and N15 respectively, and gamma distributed base measure with standard deviations one, eight, and fifteen, g1, g8, and g15 respectively) and its affect on inferences. The six subplots correspond to coefficients for each of the six covariates in Table 4 and present posterior mean (denoted by an 'x'), posterior standard deviation (delineated by the error bars being two standard deviations from the mean), and the 2.5th and 97.5th percentiles (denoted by circles '○').

Table 1

Estimates of $\boldsymbol{\beta}$ from 50 simulation data sets under each frailty distribution. True $\boldsymbol{\beta} = (1, 0.69)$ with $n = 200$ subjects, $K_i = 2$ observations each.

	Normal frailty	Bimodal frailty
$\hat{\beta}_1$ (sd)	0.99 (0.16)	1.00 (0.13)
$\hat{\beta}_2$ (sd)	0.71 (0.13)	0.67 (0.09)

Table 2

Baseline predictor variables included in analyses. Family history of glaucoma and the three systemic variables are patient-specific; all other variables are eye-specific. Last column presents estimates of the posterior probability the corresponding regression coefficient is non-zero in the model selection routine.

Variable	Units	Abbreviation	Post prob
Visual field variables (based on SAP)			
1. Mean deviation	decibels (dB)	MD	0.998
2. Short-term fluctuation	dB	SF	0.022
3. Pattern standard deviation	dB	PSD	0.018
4. Corrected PSD	dB	CPSD	0.001
Clinical variables			
5. Intraocular pressure	mmHg	IOP	1.000 *
6. Family history of glaucoma	factor	Hx	0.016
7. Vertical CD	no units	VCD	1.000 *
8. Shortwave automated perimetry result	factor	BLSWAP	1.000 *
Systemic/Demographic variables			
9. Race (African-Am vs. all others)	factor	RACE	0.028
10. Gender	factor	GENDER	1.000
11. Age at baseline	years	AGE	1.000

* Note that BLSWAP, IOP, and VCD are forced into the model thus effectively receiving a posterior probability of one.

Table 3

Posterior probability of model selected by the Hastings sampler. Variable number corresponds to the listing in Table 2. In the model selection routine, IOP (variable #5), VCD (#7), and BLSWAP (#8) are forced into the model. The posterior probability of visits to the models not listed was less than 0.1%. The table is read as all models presented include the variables numbered 5,7,8,10,11 plus the variables listed in the second row.

Models	5,7,8,10,11 +							
		1	1,9	1,6	1,2	1,3	1,3,9	1,3,6
Posterior prob	0.013	0.906	0.027	0.014	0.022	0.015	0.001	0.002

Table 4

Descriptive summary of the posterior distributions of $\boldsymbol{\beta}$ from the SWAP data.

	Mean	Std dev	median	2.5 percentile	97.5 percentile
Gender	−2.0379	0.6031	−2.0222	−3.3016	−0.9011
Age	0.1206	0.0201	0.1206	0.0811	0.1600
IOP	0.1960	0.0525	0.1963	0.0914	0.2991
VCD	3.8445	1.1801	3.8346	1.5508	6.2004
MD	−0.7096	0.2191	−0.7179	−0.1.1062	−0.2731
BLSWAP	0.9688	0.5204	0.9747	−0.0663	1.9618