

Now BU 813

BU-1442-M

A LATENT CLASS MIXED MODEL FOR ANALYZING BIOMARKER TRAJECTORIES
WITH IRREGULARLY SCHEDULED OBSERVATIONS

HAIQUN LIN,¹ CHARLES E. MCCULLOCH,¹ BRUCE W. TURNBULL,^{1*} ELIZABETH
H. SLATE,¹ AND LARRY C. CLARK²

¹Department of Statistical Science, Cornell University,
Ithaca, NY 14853

²Epidemiology Program, Arizona Cancer Center, University of Arizona,
Tucson, AZ 85716

* Correspondence to: Bruce Turnbull, Department of Operations Research and Industrial
Engineering, 227 Rhodes Hall, Cornell University, Ithaca, NY 14853

Addresses, phone numbers and emails of the authors:

Haiqun Lin, 443 Warren Hall, Cornell University, Ithaca, NY 14853.

Phone number: 607-255-2639. Email : HL18@cornell.edu

Charles E. McCulloch, 439 Warren Hall, Cornell University, Ithaca, NY 14853.

Phone number: 607-255-1643. Email : cem1@cornell.edu

Bruce W. Turnbull, 227 Rhodes Hall, Cornell University, Ithaca, NY 14853.

Phone number: 607-255-9131. Email : bruce@orie.cornell.edu

Elizabeth H. Slate, 228 Rhodes Hall, Cornell University, Ithaca, NY 14853.

Phone number: 607-255-9131. Email : slate@orie.cornell.edu

Larry C. Clark, Epidemiology Program, Arizona Cancer Center, University of Arizona,
Tucson, AZ 85716.

Phone number: 520-321-7798. Email: LCCClark@u.arizona.edu

A LATENT CLASS MIXED MODEL FOR ANALYZING BIOMARKER TRAJECTORIES WITH IRREGULARLY SCHEDULED OBSERVATIONS

SUMMARY

This paper considers a latent class model to uncover subpopulation structure for both biomarker trajectories and the probability of disease outcome in highly unbalanced longitudinal data. A specific pattern of trajectories can be viewed as a latent class in a finite mixture. Membership in latent classes is modeled with a polychotomous logistic regression. The biomarker trajectories within a latent class are described by a linear mixed model with possibly time-dependent covariates and the probabilities of disease outcome are estimated via a class specific model. Thus the method characterizes biomarker trajectory patterns to unveil the relationship between trajectories and outcomes of disease. The coefficients for the model are estimated via a generalized EM algorithm (GEM), a natural tool to use when latent classes and random coefficients are present. Standard errors of the coefficients are calculated by bootstrapping. The model fitting procedure is illustrated with data from the Nutritional Prevention of Cancer trials; we use prostate specific antigen (PSA) as the biomarker for prostate cancer and the goal is to examine trajectories of PSA serial readings in individual subjects in connection with incidence of prostate cancer.

1. INTRODUCTION

We consider the situation where a group of individuals has been followed for a period of time for incidence of a specific disease of interest. At irregular intervals during this period, biomarker readings, measured on a continuous scale, were taken from each subject. The goal is to use information on disease outcomes, the patterns of serial biomarker readings available for the subjects, together with covariates, fixed and time-varying, to discover underlying subpopulation structure. The identified subpopulations further our understanding of the relationships among the biomarker readings, the covariates and the disease outcomes and can form the basis of a prognostic tool for classifying patients.

Our work was motivated by data from the Nutritional Prevention of Cancer (NPC) trials¹. Those two trials, the first of which began in 1983, followed 1,736 subjects, some for ten years or more. The participants had been randomized into two equal groups -- half receiving a daily nutritional supplement of selenium (Se), the others placebo. Many endpoints were recorded, and we shall analyze the incidence of prostate cancer among the 1,229 men in the trial. At approximate six-month intervals, blood samples were taken from each subject; however, the interval lengths were highly variable. These blood samples were frozen, but later were thawed and analyzed retrospectively for the level of prostate specific antigen (PSA). More details of the data set will be presented in Section 4.

The value of PSA as a biomarker for prostate disease has been documented². Further testing for the suspicion of the presence of a tumor in the prostate is often recommended when a single reading of PSA exceeds a cutoff of 4 ng/ml. However, because of between patient variability, the sensitivity and specificity of this procedure is poor, leading to many false positives and false negatives². However recent studies have shown that use of a series of PSA readings over time can lead to more accurate diagnoses³. Longitudinal biomarkers are important for other diseases, such as CD4 T-cell counts for onset of clinical AIDS in HIV+ patients and AFP (alpha fetoprotein) for hepatocellular carcinoma in hepatitis B carriers and patients with liver cirrhosis. Emir et al⁴ have looked at CEA (carcinoembryonic antigen), CA15-3 (a MUC1 mucin gene product) and TPS (tissue polypeptide specific antigen) as longitudinal markers for the progression of breast cancer. We use the methods developed in

this paper to expose subpopulation structure depending on patterns of PSA readings, prostate cancer incidence, and Se and other covariates among the NPC participants.

Logistic regression can be used to classify individuals when there are known subgroups⁵, but it is not designed for handling longitudinal data. In any case, the subgroups are not predefined in our application. Regression tree methodology, such as CART⁶ is a standard nonparametric method that can handle unknown subgroups. However, when applying CART to longitudinal data⁷, there are difficulties accommodating time-varying covariates and unequally spaced and highly unbalanced biomarker readings. In addition, such methods do not recognize the distinct roles that biomarker trajectories, covariates and disease outcome variables play. The same problems afflict the multivariate adaptive regression splines approach for longitudinal data (MASAL) of Zhang⁸. Therefore a parametric modeling approach may be advantageous.

An established method for longitudinal data is the two-stage random effect model of Laird and Ware⁹. This approach does not permit distinct subpopulation structure, which is one of our primary interests. Belin and Rubin¹⁰ and Rubin and Wu¹¹ used a generalized finite mixture model for pre-determined distinct subpopulations to analyze repeated measurement data. Peng, Jacobs, and Tanner^{12,13} used finite mixtures-of-experts models to estimate both the parameters in the component mixture models and the mixture fractions for pattern classification. Qu, Tan, and Kutner¹⁴, Hadgu and Qu¹⁵ and Yang and Becker¹⁶ incorporated latent classes in a finite mixture model to uncover the structure of subpopulations. However, none of these papers additionally incorporate the outcome data into patterns of longitudinal response. In a recent interesting paper, Muthen and Shedden¹⁷ do include these features. We will generalize their approach to accommodate irregularly spaced longitudinal readings and unequal numbers of readings for each subject. The model we present here can also be viewed as a generalization of the “mixtures-of-experts” models. Moreover we shall illustrate the utility of this model for biomarker and outcome data, specifically obtaining interesting results for the PSA and prostate cancer incidence data from the NPC trial.

In our analysis, we model the association between the longitudinal PSA marker process and the outcome of prostate cancer. We discuss the role of selenium and other covariates. Three subpopulations emerge from our analysis with interpretable identities -- details are

given in Section 4. This application illustrates a flexible, descriptive methodology that is suitable for longitudinal measurements in many situations.

2. THE LATENT CLASS MIXED MODEL

2.1. General Description of the Model

The search for subpopulation structure leads us to a model that contains K latent classes, with each class representing a subpopulation. We model the probability that a subject belongs to a latent class with a polychotomous logistic regression, which we allow to depend on covariates, x_* say, specific to the subject. Each subpopulation has its own model for both the biomarker readings, y , and the disease outcome, w . Conditional on a subject being a member of a subpopulation, we model the continuous biomarker readings using a Laird and Ware⁹ style mixed model into which we incorporate time-varying and fixed covariates, x , and subject-specific random effects, u , modeled through the design matrix for the random effects, z . The probability of disease outcome is also allowed to be latent class specific in order to further capture subpopulation structure. In the following subsections we give details of each portion of the model.

2.2. Modeling the Probability of Latent Class Membership

Let $c_i^* = (c_{i1}, \dots, c_{iK})'$ have a multinomial distribution, in which $i = 1, \dots, n$ is the index for subjects, $k = 1, \dots, K$ is the index for latent classes, and c_{ik} is an indicator of class k for subject i . The categorical variables in c_i^* are modeled as a function of covariates $x_{*i} = (x_{*i1}, \dots, x_{*im})'$, which would usually include an intercept term, in a multinomial logistic regression. The probability that subject i falls into class k is:

$$\pi_{ik} := P\{c_{ik} = 1\} = \frac{\exp(x_{*i}'\alpha_k)}{\sum_{j=1}^K \exp(x_{*i}'\alpha_j)}, \quad (1)$$

where α_k is the vector containing the coefficients for class k and $\alpha_K = 0$.

2.3. Modeling Trajectories of the Longitudinal Biomarker

We have repeated measurements for the continuous biomarker readings, y , which are modeled as a linear function of fixed and random effects as:

$$y_i = x_i \beta + z_i u_i + \varepsilon_i. \quad (2)$$

Here, $y_i = (y_{i1}, \dots, y_{iT_i})'$, and T_i is the number of biomarker readings for the i th subject. The $T_i \times p$ matrix of covariates is denoted $x_i' = (x_{i1}, \dots, x_{iT_i})'$, and will typically include time and functions of time. The t th row x_{it} in x_i is the t th vector of covariate values for subject i . Covariates for random effects are in $z_i' = (z_{i1}, \dots, z_{iT_i})'$, which is a $T_i \times q$ sub-matrix of x_i with $q \leq p$. Here β and u_i are p and q -dimensional vectors of fixed and random coefficients respectively. The error ε_i is a T_i -dimensional vector uncorrelated with other variables, multi-normally distributed with mean 0 and covariance matrix $\sigma^2 I_{T_i}$.

The biomarker readings, y_i are related to the latent class through the random coefficient u_i whose mean is class specific. In many longitudinal studies, data are irregularly timed and the number of measurements differs among subjects. It is therefore usually difficult to model the covariance structure of y_i . However, using latent classes and random effects, along with independent errors, is a natural way to specify a complicated covariance structure.

2.4. Modeling the Mean of the Random Coefficients

The random coefficients u_i are assumed to have a distribution whose mean is determined by latent class membership through the equation:

$$u_i = \Lambda c_i + \zeta_i. \quad (3)$$

Here Λ is a $q \times (K-1)$ matrix of coefficients, $\Lambda = (\lambda_1, \dots, \lambda_k, \dots, \lambda_{K-1})$, and λ_k is a q -dimensional column vector containing the means of the random coefficient u_i for latent class k . Given class membership for subject i being k , i.e. $c_{ik} = 1$, we have $E(u_i) = \lambda_k$ for $k=1, \dots, K-1$ and $E(u_i) = 0$ if subject i is in class K . The vector c_i has the first $K-1$ elements of c_i^* defined in (1). Also ζ_i is a q -dimensional vector of errors uncorrelated with other variables, multi-normally

distributed with mean 0 and common covariance matrix Ψ across classes.

The model in (2) and (3) captures common characteristics of PSA trajectories within a subpopulation through latent classes while accommodating the variability among subjects in the same subpopulation through random effects.

2.5. Modeling the Binary Outcome for a Given Subject

Let w_i denote the binary variable for disease outcome for subject i . It is an indicator of whether the i th subject experiences a certain outcome. We assume a class-specific probability for disease outcome through binary logistic regression with intercept only as:

$$g_k := P\{w_i = 1 \mid c_{ik} = 1\} = \frac{\exp(\rho_k)}{1 + \exp(\rho_k)}. \quad (4)$$

Here ρ_k is the log-odds of the occurrence of the outcome for a member of class k .

2.6. Comments on the Above Model

The relations in (2) and (3) represent a mixture of normal mixed effects for the continuous marker variable y_i . Given latent class membership c_i , y_i and w_i are assumed independent; this is an important feature of the latent class model and greatly simplifies the modeling procedure. However, y_i and w_i are still marginally correlated since the probability of the outcome is class-specific as in (4). Subjects within the same latent class are also correlated since their random coefficients come from the same distribution as in (3).

3. MODEL FITTING

The maximum likelihood method is used to obtain coefficient estimates in the above model. Since likelihood equations of the observed data (y , w and x) have very complicated expressions, we therefore will use the EM algorithm¹⁸ to obtain parameter estimates. Standard errors of the estimates are calculated via the bootstrap method.

3.1. Parameter Estimation via the GEM Algorithm

We view the random coefficient u_i 's and the latent class variable c_i 's as missing data. Let $[A|B]$ denote the conditional probability density of A given B ; then the complete-data log-likelihood, $\log L_c$, can be written as:

$$\log L_c = \sum_{i=1}^n (\log[c_i | x_i] + \log[u_i | x_i, c_i] + \log[w_i | x_i, c_i] + \log[y_i | x_i, u_i]), \quad (5)$$

$$\text{where } \sum_{i=1}^n \log[c_i | x_i] = \sum_{i=1}^n \sum_{k=1}^K c_{ik} \log \pi_{ik} = \sum_{i=1}^n \sum_{k=1}^K c_{ik} x_{*i} \alpha_k - \sum_{i=1}^n \log \left(\sum_{j=1}^K \exp(x_{*i} \alpha_j) \right); \quad (6)$$

$$\sum_{i=1}^n \log[u_i | x_i, c_i] = -\frac{nq}{2} \log 2\pi + \frac{n}{2} \log |\Psi^{-1}| - \frac{1}{2} \text{tr} \left(\sum_{i=1}^n \Psi^{-1} (u_i u_i' + \Lambda c_i c_i' \Lambda' - 2u_i c_i' \Lambda') \right); \quad (7)$$

$$\begin{aligned} \sum_{i=1}^n \log[w_i | c_i] &= \sum_{i=1}^n \sum_{k=1}^K c_{ik} (w_i \log g_k + (1-w_i) \log(1-g_k)) \\ &= \sum_{i=1}^n \sum_{k=1}^K c_{ik} (w_i \rho_k - \log(1 + \exp(\rho_k))); \end{aligned} \quad (8)$$

$$\begin{aligned} \sum_{i=1}^n \log[y_i | x_i, u_i] &= \sum_{i=1}^n \left(-\frac{T_i}{2} \log(2\pi) - \frac{T_i}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y_i - x_i \beta - z_i u_i)' (y_i - x_i \beta - z_i u_i) \right) \\ &= -\frac{1}{2} \sum_{i=1}^n T_i \log(2\pi) - \frac{1}{2} \sum_{i=1}^n T_i \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i' y_i - y_i' x_i \beta - y_i' z_i u_i \\ &\quad - \beta' x_i y_i + \beta' x_i x_i \beta + \beta' x_i z_i u_i - u_i' z_i y_i + u_i' z_i x_i \beta + u_i' z_i z_i u_i). \end{aligned} \quad (9)$$

Conditioning on the observed data, y_i and w_i , expectations of the terms involving c_i and u_i in (6) – (9) are calculated in the E-step. Parameter estimates are obtained by maximizing the conditional expectations in M-step. One step of the Newton-Raphson method with adjusted step length for each iteration in the M-step¹⁸ is used to estimate α and ρ in (6) and (8) because the likelihood equations for the two logistic regressions are non-linear and do not have closed form maximizers. Hence the algorithm is a generalized EM (GEM), see Appendix A.3.1. For Λ , ψ , β and σ^2 in (7) and (9), estimates are obtained in standard fashion. The E- and M-steps are repeated until differences in values of all the estimates in the consecutive iterations are sufficiently small. We used a convergence criterion of 10^{-4} . Detailed calculations of the E- and M-steps are described in the Appendix. Our GEM algorithm was implemented in Matlab¹⁹. Multiple starting values of the parameters were used for the GEM algorithm and we chose the solution yielding the highest log-likelihood value for a given number of latent classes.

3.2. Calculation of Standard Errors for the Estimates via Parametric Bootstrap

Using our model assumptions, marker readings y and binary outcome variables w for a parametric bootstrap sample were simulated using the final estimates from the GEM algorithm and covariates x for all the subjects in the original data. The model was then refitted to each bootstrap sample by the above GEM algorithm. A total of 200 bootstrap samples were so generated, from which standard errors of our parameter estimates were obtained.

4. APPLICATION TO THE PSA DATA SET

4.1. Description of the Data Set and the Variables

PSA is a small protein secreted into the blood by prostate tissue, both normal and malignant. The concentration of PSA in the blood varies with the amount of normal prostate tissue, the amount of cancer, the location and type of cancer, and the extent of any existing infection or inflammation of the prostate. A normal blood PSA value is usually less than 3-4 ng/ml, and may vary with age, but about 20% of men diagnosed with prostate cancer have measured PSA levels less than 4 ng/ml²⁰. With an elevated PSA reading, an ultrasound, biopsy or other test may be recommended by physicians to further evaluate the prostate. However, because other conditions such as benign prostatic hyperplasia or prostatitis also can increase PSA levels, a single high PSA measurement is not a reliable indicator of prostate cancer. Recently, interest has focused on the utility of a series of PSA readings for detecting prostate cancer and determining the appropriate medical attention.

We use the series of PSA readings obtained from the frozen blood samples of the NPC participants and the corresponding prostate cancer outcomes to illustrate our latent class mixed model. Recall that in this trial the subjects were randomly assigned to receive daily either placebo or a nutritional supplement of Se. Thus we seek to use our methodology to discover latent structure in the PSA trajectories and prostate cancer outcomes, depending on the Se status of the NPC subjects. The data available consist of treatment assignment (Se or placebo), prostate cancer diagnosis date (if any), baseline PSA and Se blood levels, and

longitudinal PSA and Se readings recorded at approximate six-month intervals for 1229 men. We restrict the data used for our analyses in two ways. First, we determine the outcome (prostate cancer or not) for a subject at 7 years after randomization. Thus we include only those observations within 7 years of randomization and before prostate cancer diagnosis. This 7-year window was chosen to balance the number of longitudinal readings available and the number of subjects for whom follow up was complete. Whittemore et al.²¹ also used a 7-year window in their analyses. Second, we omit observations for which the baseline Se level is missing since Se is used as a covariate both in predicting class-membership and PSA levels in our model. The resulting data set that we analyze consists of 979 subjects with a highly variable number of measurements (range 1-16, median 3) at irregularly spaced intervals. Of these 979 men, 74 were not followed for at least seven years, among whom, only 12 were not followed for six years. All patients had been followed for more than 5 years and 4 months.

The variables we include in our model are as follows: the covariate vector x_* used to predict latent class membership in (1) contains the indicator variable of Se supplementation, and the values of age, PSA and Se level at randomization. The longitudinal marker value y in (2) is the vector of PSA readings transformed to $\log(PSA+1)$; this transformation has also been used by Pearson et al.²² and Whittemore et al.²¹. The fixed effects covariate vector x in equation (2) contains the indicator of Se supplementation, the value of age, and Se level at randomization, and linear and quadratic terms of visit time expressed in years since entry into the trial. The covariates for the random effects in (2), z , also contain years since entry up to quadratic terms. Finally, the outcome variable w in (4) is an indicator of prostate cancer diagnosis within 7 years of entry. In our data, 49% of subjects are in the Se supplementation group. As shown in Table 1, the two treatment groups have similar baseline average values of age, Se and PSA. Also, examination of the longitudinal PSA readings reveals little difference between the two groups. Not shown, but as expected, the longitudinal Se levels in the Se supplementation group rapidly diverge from those in the placebo group, stabilizing at a higher level of approximately 164 ng/ml while the average Se levels in placebo group remains almost the same as that at randomization.

Table 1: Average Covariate Values (range) by Treatment Group

Baseline Variable	Placebo group (N=500)	Se supplement group (N=479)
Age at entry, years	63.8 (34.0, 80.7)	64.7 (18.9, 80.7)
Se level at entry, ng/ml	114.0 (47.6, 209.2)	115.7 (55.0, 195.6)
PSA level at entry, ng/ml	2.1 (0.0, 69.0)	1.8 (0.0, 167.4)

4.2. Fitting the Model to the Data

We start from the relatively simple case of a single latent class. Then the model reduces to a standard linear mixed model. For models with two or more classes, we proceed as in Section 3. The maximized log-likelihood values corresponding to the $K=1, 2, 3, 4$ and 5 latent class models are listed in Table 2. The maximized log-likelihood of the one class model is -1356.2 --low compared to that of the models with latent classes. Using the deviances from Table 2, $K=4$ appears to be the favored number of latent classes; $K=5$ yields a non-significant improvement.

Two of the PSA trajectories identified by the four-class model are almost the same in terms of trajectory shape (Figure 1) and prostate cancer outcome (Table 3). The three-class solution, on the other hand, identifies three distinct PSA trajectories. The observed incidence of disease is closely matched by both the fit of the three- and four-class solutions. And residual plots for the biomarker for both three- and four-class models show adequate fit. Details of obtaining these fits and residuals are given in Section 4.4. We report detailed calculations for the four-class model.

Table 2: Log-Likelihood by Number of Latent Classes, K

K	Maximized log-likelihood	Deviance change*	Difference in number of parameters*
1	-1356.2	1513.4	36
2	-912.7	626.4	27
3	-672.2	145.4	18
4	-600.0	1.0	9
5	-599.5	-	-

* these values are measured by comparison with the 5-class model

The four-class solution identifies trajectories classes that we shall label as “*Low1*”, “*Low2*”, “*Medium*” and “*High*”. Figure 1 shows the fitted biomarker values at half year

intervals, calculated for each class using the formula $\log(1 + \hat{PSA}) = x\hat{\beta} + z\hat{\lambda}_k$ and inserting average values for each covariate. PSA was transformed back to the original scale for plotting.

The “*Low1*” and “*Low2*” classes are characterized by a consistently low PSA level (0.6 and 1.5 ng/ml respectively) throughout the trial period. The estimated percentage of subjects in the two “*Low*” classes are 28% and 46% respectively. Table 3 shows that the “*Low1*” and “*Low2*” classes both have the lowest estimated probability of developing prostate cancer (0.1% and 0.9%). The “*Medium*” class has a higher PSA level than that of the two “*Low*” classes throughout the trial, as evidenced in Figure 1. The PSA level increases slightly over time for this class and the estimated probability of developing prostate cancer for patients in this class (10.9%) is between that for those in the “*Low*” and the “*High*” classes. The “*High*” class has the highest PSA level at the beginning of the trial, and the predicted level of PSA increases over time. About 56% of members in this class are estimated to develop prostate cancer within 7 years (Table 3). For comparison, in the three-class solution, we designate the three groups as “*Low*”, “*Medium*” and “*High*” (Figure 1, lower panel). The proportion in the “*Low*” class is 65%, which closely matches the proportion of the two “*Low*” classes in the four-class solution. Also, the “*Medium*” and “*High*” classes are similar between the three- and four-class solutions.

The estimated incidence probabilities, \hat{g}_k , are calculated by plugging the maximum likelihood estimates $\hat{\rho}_k$ into equation (4). Similarly, class membership probabilities $\hat{\pi}_{ik}$ are estimated from (1). Each of the four classes have $\hat{\pi}_{ik}$ ’s that are nearly one for some of the subjects, implying that all the four classes can be identified.

Tables 3 through 7 list parameter estimates for the four-class model; further discussion of the significance of effects is in Section 4.5. In Tables 4 and 6 we use “*Low2*” as the baseline class since it is estimated to contain the largest proportion of subjects – almost half. To compare the observed PSA trajectories with the fitted ones in Figure 1, we proportionally allocated subjects to the classes using the $\hat{\pi}_{ik}$ ’s. An observed trajectory for each class was calculated by a weighted average using the allocated proportions. Figure 2 shows that we have four distinct observed trajectories. The trajectory for the “*High*” class in Figure 2 differs from that in Figure 1 because of selection bias: the contributions of the many observations with

small weights from the other classes become increasingly dominant over time as subjects from the “High” class are diagnosed with prostate cancer and are dropped from the analysis.

Table 3: Estimated Probability for the Binary Outcome (\hat{g}) & Estimated Class Proportions

Class	\hat{g} (95% C.I.)	$\hat{\rho}$ (std. err.)	Class proportion
“Low1”	0.0001 (0.0000, 0.0013)	-9.2835 (1.3396)	0.2731
“Low2”	0.0087 (0.0025, 0.0301)	- 4.7350 (0.6433)	0.4606
“Medium”	0.1092 (0.0726, 0.1610)	-2.0993 (0.2288)	0.2256
“High”	0.5566 (0.3845, 0.7161)	0.2273 (0.3561)	0.0407

¹The deviance change for testing that probabilities of developing prostate cancer are the same for all classes is 137.51 over 2 d.f.

Table 4: Estimated α_k for Class Membership

Covariate	Deviance change	Class	Estimate	Std. Err.
<i>Intercept</i>	874.43	“Low1”	23.5698	(11.23)
		“Medium”	-25.8175	(0.739)
		“High”	-55.0802	(15.48)
<i>Treatment group</i>	4.00	“Low1”	-2.2426	(1.734)
		“Medium”	-0.2797	(0.569)
		“High”	-1.9284	(1.359)
<i>Age at entry</i>	0.89	“Low1”	-0.0607	(0.486)
		“Medium”	0.7685	(0.346)
		“High”	0.3275	(0.847)
<i>log(PSA+1) at entry</i>	2326.46	“Low1”	-45.6412	(22.01)
		“Medium”	22.6165	(0.390)
		“High”	37.3498	(7.857)
<i>Se at entry</i>	0.50	“Low1”	-5.5612	(4.480)
		“Medium”	1.9841	(0.268)
		“High”	3.6535	(2.733)

Note: “Low2” is fitted as the baseline class, K.

Table 5: Estimates for the Fixed Effect Coefficients (β) and σ^2

Covariate	Estimates	(Std.err.)	Deviance change
<i>Intercept</i>	0.7493	(0.0241)	1665.82
<i>Treatment group (Se supplementation)</i>	-0.0264	(0.0173)	0.51
<i>Age at entry</i>	0.0152	(0.0088)	0.80
<i>Year since entry (linear)</i>	0.0491	(0.0123)	33.70
<i>Year since entry (quadratic)</i>	-0.0039	(0.0015)	12.23
<i>Se at entry</i>	-0.0879	(0.0126)	4.74
Estimated σ^2	0.0450	(0.0001)	

Table 6: Class-specific Mean Difference of Random Effects Λ from the “Low2” Class

Trend	Deviance change	Class	Estimate	Std.err.
Intercept	2894.7	“Low1”	-0.4209	(0.0334)
		“Medium”	0.6783	(0.0335)
		“High”	1.7253	(0.0615)
Linear	399.9	“Low1”	-0.0137	(0.0129)
		“Medium”	-0.0135	(0.0127)
		“High”	0.0571	(0.0248)
Quadratic	4.4	“Low1”	0.0006	(0.0010)
		“Medium”	0.0027	(0.0011)
		“High”	-0.0011	(0.0020)

Table 7: Estimates for the Covariance Matrix Ψ of Random Coefficient Means (Bootstrapping standard errors in parenthesis)

	<i>Intercept</i>	<i>Linear</i>	<i>Quadratic</i>
<i>Intercept</i>	0.0442 (0.0096)	-0.0045 (0.0051)	-0.43e-5 (0.0001)
<i>Linear</i>	-0.0045 (0.0051)	0.0037 (0.0040)	-0.80e-4 (0.0005)
<i>Quadratic</i>	-0.43e-5 (0.0001)	-0.80e-4 (0.0005)	0.27e-5 (0.48e-4)

4.3. Characterization of the Latent Classes

In Section 4.2, we described the disease marker trajectories for each latent class. We now turn to a description of the latent classes themselves, through both covariate values and marker trajectories. Table 8 summarizes covariate values by latent classes, revealing some interesting features for covariates in each class. The “High” class has a remarkably low proportion in the Se supplementation group (37%), while the percentage of subjects in the Se supplementation group is 49% in the study population. Because this class has the highest estimated probability of prostate cancer, this observation supports the hypothesized beneficial effect of Se supplementation¹. Also, it is noticed that subjects in the “Medium” and “High” classes have much higher average age at entry than those in the two “Low” classes. All classes have similar baseline Se levels.

Table 8: Average Covariate Values by Latent Classes

Baseline Variables	Latent Classes			
	"Low1"	"Low2"	"Medium"	"High"
Age at entry, years	62.0	63.4	67.8	68.7
Se level at entry, ng/ml	107.6	117.6	117.6	115.8
PSA level at entry, ng/ml	0.42	1.21	3.42	12.54
Proportion in Se. group	0.45	0.51	0.53	0.37

4.4. Assessing Fit of the Model

Marginal predicted values can be obtained for the observed responses y and w .

The marginal distribution of y_i is a normal mixture with density given by:

$$[y_i | x_i] = \sum_{k=1}^K [y_i | x_i, c_{ik} = 1][c_{ik} = 1 | x_i] = \sum_{k=1}^K N_{T_i}(x_i \beta + z_i \lambda_k, z_i \Psi z_i' + \Sigma_i) \pi_{ik}. \quad (10)$$

The predicted y_i 's (fitted $\log(\text{PSA}+1)$) are calculated as estimated weighted means according to (10). The residuals of $\log(\text{PSA}+1)$ versus year are plotted as shown in Figure 3. It can be seen that almost all residuals fall within 2 units around 0 indicating adequate fit.

The marginal distribution of w can be calculated as:

$$P\{w_i = 1\} = \sum_{k=1}^K [y_i | x_i, c_{ik} = 1][c_{ik} = 1 | x_i] = \sum_{k=1}^K g_k \pi_{ik}. \quad (11)$$

and thus the marginal fitted value for w can be obtained via: $\hat{P}\{w_i = 1\} = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \hat{g}_k \hat{\pi}_{ik}$. (12)

The fitted value of $P\{w_i = 1\}$ for the 4-class solution is 0.0518, which matches closely the observed proportion of prostate cancer 0.0531.

4.5. Discussion of Significance of Effects for the Model

The deviance approach is used to assess the significance of a given effect for our 4-class solution. The deviance change is twice the difference in the maximized log-likelihood between models with and without inclusion of the effect parameters to be assessed. The deviance changes are in Tables 3 through 7, alongside the parameter estimates.

Among those predictors in x_* used for the class membership model (1), Table 4 reveals that PSA at entry is significant with estimated coefficients in monotone agreement with the group ordering.

We turn our attention now to the fixed effects for the biomarker trajectory model (2). It can be seen from Table 5 that the intercept and time since entry are significant. Again Se supplementation is not significant, but is associated with lower PSA readings. And higher baseline Se level is significantly associated with lower PSA readings.

Examination of the deviances for the class-specific means of the random coefficients in model (3) reveals that the intercepts and terms of time since entry all vary significantly across the latent classes. This means that the PSA level at entry is quite different, as are the patterns of PSA readings over time among the four classes (Table 6). The estimated intercepts for the “*Low1*”, “*Low2*”, “*Medium*” and “*High*” classes are 0.33, 0.75, 1.43 and 2.47 respectively on the log scale (Tables 5 and 6) with an estimated within-class standard deviation of 0.21 ($\hat{\Psi}_{11}^{1/2}$ in Table 7).

Finally, testing the hypothesis that the probability of developing prostate cancer is the same for all three classes results in a change in deviance of 137.51 with 2 fewer parameters. We infer that chances for the prostate cancer differ greatly among the four classes (Table 3).

6. Concluding Remarks

Fitting the model to the PSA data illustrates the explanatory power and flexibility of the latent class mixed model. The model fits the data well and is able to uncover subpopulation structure. We discussed mainly the four-class solution which identified “*Low1*”, “*Low2*”, “*Medium*” and “*High*” as statistically significant latent classes; however, for many practical purposes, the three-class model may be satisfactory. For this model the PSA trajectory patterns are clearly distinct among different classes, as are the estimated probabilities of prostate cancer. Since the “*High*” class has the highest probability of developing prostate cancer, the fact that the proportion of patients in the selenium supplement group for the “*High*” class is dramatically lower than other classes supports the hypothesis that selenium supplementation is beneficial in prostate cancer prevention¹. Sequential estimation of the π_{ik} could be a valuable prognostic tool for classifying subjects. Similar dynamic updating ideas

have been developed for simpler longitudinal biomarker models^{23, 24, 25}. Our latent class results suggest ways to form prognostic indicators for assessing the risk of prostate cancer for individuals. The parametric nature of the model provides explanatory power for the estimates and correlation in longitudinal data is more easily taken into account.

APPENDIX: COMPUTATIONAL FORMULAE

A.1. The Observed-data Log-likelihood

By the conditional independence of w and y , the observed-data log-likelihood $\log L_o$, may be

$$\text{expressed as: } \sum_{i=1}^n \log[y_i, w_i | x_i] = \sum_{i=1}^n \log \left(\sum_{k=1}^K [y_i | x_i, c_{ik} = 1][w_i | x_i, c_{ik} = 1][c_{ik} = 1] \right),$$

$$\text{since } [y_i | x_i, c_i] = N_{T_i}(x_i\beta + z_i\Lambda c_i, z_i\Psi z_i' + \Sigma_i). \quad (13)$$

$$\text{The log } L_o \text{ becomes: } \sum_{i=1}^n \log \left(\sum_{k=1}^K N_{T_i}(x_i\beta + z_i\lambda_k, z_i\Psi z_i' + \Sigma_i) g_k^{w_i} (1 - g_k)^{1-w_i} \pi_{ik} \right). \quad (14)$$

A.2. The E Step

We need to calculate the conditional expectations of c_{ik} , $u_i u_i'$, $c_i c_i'$ and $u_i c_i'$ in (6), (7) and (8), and the conditional expectation of u_i and $u_i' z_i z_i' u_i$ in (9).

$$\begin{aligned} \text{Let } p_{ik} &:= E(c_{ik} | x_i, y_i, w_i) = \frac{[c_{ik} = 1 | x_i][y_i, w_i | x_i, c_{ik} = 1]}{[y_i, w_i | x_i]} \\ &= \frac{\pi_{ik}[y_i | x_i, c_{ik} = 1][w_i | x_i, c_{ik} = 1]}{\sum_{k=1}^K \pi_{ik}[y_i | x_i, c_{ik} = 1][w_i | x_i, c_{ik} = 1]}. \end{aligned} \quad (15)$$

Again, by conditional independence of w and y , and let $p_i = (p_{i1}, \dots, p_{iK-1})'$. Expectation of

$$c_i c_i' \text{ can be calculated as: } \tilde{E}_{cc} := E\left(\frac{1}{n} \sum_{i=1}^n c_i c_i' | x_i, y_i, w_i\right) = \frac{1}{n} \sum_{i=1}^n \text{diag}(p_i). \quad (16)$$

Next, $E(u_i c_i' | x_i, y_i, w_i)$ needs to be calculated. However, First note that

$$\text{Cov}(y_i, u_i) = \text{Cov}(x_i \beta + z_i u_i + \varepsilon_i, u_i) = z_i E u_i u_i' + E \varepsilon_i u_i' - z_i E u_i E u_i' = z_i \Psi. \quad (17)$$

And by the properties of conditional probability for bivariate normal variables, we have:

$$E(u_i | y_i, c_i, x_i) = \Lambda c_i + \Psi z_i' (z_i' \Psi z_i + \Sigma_i)^{-1} (y_i - x_i \beta - z_i \Lambda c_i) = V_i (\Psi^{-1} \Lambda c_i + z_i' \Sigma_i^{-1} (y_i - x_i \beta)), \quad (18)$$

$$\text{where } V_i = \text{Cov}(u_i, u_i | y_i, c_i, x_i) = (z_i' \Sigma_i^{-1} z_i + \Psi^{-1})^{-1}. \quad (19)$$

$$\begin{aligned} \text{Then } \tilde{E}_{uc} &:= \frac{1}{n} \sum_{i=1}^n E(u_i c_i' | x_i, y_i, w_i) = \frac{1}{n} \sum_{i=1}^n E_c \left((E(u_i | x_i, y_i, w_i, c_i)) c_i' \right) \\ &= \frac{1}{n} \sum_{i=1}^n V_i \left(\Psi^{-1} \Lambda \text{diag}(p_i) + z_i' \Sigma_i^{-1} (y_i - x_i \beta) p_i' \right). \end{aligned} \quad (20)$$

$E(u_i u_i' | x_i, y_i, w_i)$ can also be calculated by further conditioning on latent class:

$$E(u_i u_i' | x_i, y_i, w_i, c_i) = V_i + V_i \left(\Psi^{-1} \Lambda c_i + z_i' \Sigma_i^{-1} (y_i - x_i \beta) \right) \left(c_i' \Lambda \Psi^{-1} + (y_i - \beta' x_i)' \Sigma_i^{-1} z_i' \right) V_i. \quad (21)$$

$$\begin{aligned} \text{And } \tilde{E}_{uu} &:= \frac{1}{n} \sum_{i=1}^n E(u_i u_i' | x_i, y_i, w_i) = \frac{1}{n} \sum_{i=1}^n E_c \left(E(u_i u_i' | x_i, y_i, w_i, c_i) \right) \\ &= \frac{1}{n} \sum_{i=1}^n V_i + \frac{1}{n} \sum_{i=1}^n V_i \left(\Psi^{-1} \Lambda \text{diag}(p_i) \Lambda \Psi^{-1} + \Psi^{-1} \Lambda p_i (y_i - x_i \beta)' \Sigma_i^{-1} z_i' \right. \\ &\quad \left. + z_i' \Sigma_i^{-1} (y_i - x_i \beta) p_i' \Lambda \Psi^{-1} + z_i' \Sigma_i^{-1} (y_i - x_i \beta) (y_i - x_i \beta)' \Sigma_i^{-1} z_i' \right) V_i. \end{aligned} \quad (22)$$

To calculate the conditional expectation of (9), we note that:

$$\tilde{u}_i := E(u_i | y_i, w_i, x_i) = E_c(E(u_i | c_i, y_i, x_i)) = V_i \left(\Psi^{-1} \Lambda p_i + z_i' \Sigma_i^{-1} (y_i - x_i \beta) \right). \quad (23)$$

The last equality follows from (18). And by the properties for expectation of the quadratic form of a multivariate normally distributed vector, we have:

$$E(u_i' z_i z_i' u_i | x_i, y_i, c_i) = \text{tr}(z_i' z_i V_i) + E(u_i' | x_i, y_i, c_i) z_i' z_i E(u_i | x_i, y_i, c_i). \quad (24)$$

$$\begin{aligned} \text{Therefore, } \tilde{E}(u_i' z_i z_i' u_i) &:= E(u_i' z_i z_i' u_i | x_i, y_i, w_i) = E_c \left(E(u_i' z_i z_i' u_i | x_i, y_i, c_i) \right) \\ &= \text{tr}(z_i' z_i V + \Lambda' \Psi^{-1} V_i z_i z_i' V_i \Psi^{-1} \Lambda \text{diag}(p_i)) + (y_i - \beta' x_i)' \Sigma_i^{-1} z_i' V_i z_i \tilde{u}_i. \end{aligned} \quad (25)$$

A.3. The M Step

A.3.1 The GEM Algorithm Based on One Newton-Raphson Step

Inserting p_{ik} , the conditional expectation of c_{ik} in (15) into (6), we get:

$$E_o \left(\sum_{i=1}^n \log[c_i | x_i] \right) = \sum_{i=1}^n \sum_{k=1}^K p_{ik} \log \pi_{ik} = \sum_{i=1}^n \sum_{k=1}^K p_{ik} x_{*i}' \alpha_k - \sum_{i=1}^n \log \left(\sum_{j=1}^K \exp(x_{*i}' \alpha_j) \right). \quad (26)$$

where $E_o(\cdot)$ denote expectation conditioning on the observed data y_i and w_i . We then apply Newton-Raphson method to obtain estimate for $\alpha_k^{(k)}$. The log-likelihood of the complete data is increased by choosing $\alpha^{(k)}$ ($0 < \alpha^{(k)} \leq 1$) small enough in just one iteration of Newton-

$$\text{Raphson: } \alpha_k^{(k+1)} = \alpha_k^{(k)} + a^{(k)} \Gamma^{-1}(\alpha_k^{(k)}) S(\alpha_k^{(k)}). \quad (27)$$

This then defines a GEM sequence²⁶. Thus, we obtain estimates of α_k in $(j+1)$ th GEM

$$\text{iteration as: } \alpha_k^{(j+1)} = \alpha_k^{(j)} + a^{(j)} \left(\sum_{i=1}^n \pi_{ik}^{(j)} (1 - \pi_{ik}^{(j)}) x_{*i} x_{*i}' \right)^{-1} \left(\sum_{i=1}^n x_{*i} (p_{ik}^{(j)} - \pi_{ik}^{(j)}) \right). \quad (28)$$

Similarly, inserting p_{ik} , the conditional expectation of c_{ik} in (15) into (8), then we can easily calculate $S(\rho_k)$ and $I(\rho_k)$. Use the scheme in (27), we get:

$$\rho_k^{(j+1)} = \rho_k^{(j)} + a^{(j)} \left(\sum_{i=1}^n p_{ik} g_k^{(j)} (1 - g_k^{(j)}) \right)^{-1} \left(\sum_{i=1}^n p_{ik}^{(j)} (w_i - g_k^{(j)}) \right). \quad (29)$$

A.3.2. The M-Step for Other Parameters

Inserting \tilde{E}_{cc} , \tilde{E}_{uc} , and \tilde{E}_{uu} calculated in (17), (20) and (22) into (7), we have:

$$E_o \left(\sum_{i=1}^n \log[u_i | x_i, c_i] \right) = -\frac{nq}{2} \log 2\pi + \frac{n}{2} \log |\Psi^{-1}| - \frac{n}{2} \text{tr} \left(\Psi^{-1} (\tilde{E}_{uu} + \Lambda \tilde{E}_{cc} \Lambda' - 2\tilde{E}_{uc} \Lambda') \right). \quad (30)$$

$$\text{Then the estimates of } \Lambda \text{ and } \Psi \text{ that maximize (30) are: } \hat{\Lambda} = \tilde{E}_{uc} \tilde{E}_{cc}^{-1} \quad (31)$$

$$\text{and } \hat{\Psi} = \tilde{E}_{uu} - \tilde{E}_{uc} \tilde{E}_{cc}^{-1} \tilde{E}_{uc}' \quad (32)$$

Similarly, plugging \tilde{u}_i and $\tilde{E}(u_i' z_i z_i' u_i)$ calculated in (23) and (25) into (9), we get:

$$E_o \left(\sum_{i=1}^n \log[y_i | x_i, u_i] \right) = -\sum_{i=1}^n \frac{T_i}{2} \left(\log 2\pi - \log \sigma^2 \right) - \frac{1}{2\sigma^2} \sum_{i=1}^n \left(y_i' y_i - y_i' x_i \beta - y_i' z_i \tilde{u}_i - \beta' x_i y_i \right. \\ \left. + \beta' x_i x_i \beta + \beta' x_i z_i \tilde{u}_i - \tilde{u}_i' z_i y_i + \tilde{u}_i' z_i x_i \beta + \tilde{E}(u_i' z_i z_i' u_i) \right). \quad (33)$$

Then the estimates of β and σ^2 that maximize (30) are:

$$\hat{\beta} = \left(\sum_{i=1}^n x_i' x_i \right)^{-1} \sum_{i=1}^n x_i' (y_i - z_i \tilde{u}_i) \quad (34)$$

and

$$\hat{\sigma}^2 = (1 / \sum_{i=1}^n T_i) \sum_{i=1}^n \left((y_i - x_i \hat{\beta})' (y_i - x_i \hat{\beta} - 2z_i \tilde{u}_i) + \tilde{E}(u_i' z_i' z_i u_i) \right). \quad (35)$$

ACKNOWLEDGEMENTS

This research was funded in part by grants from the National Institutes of Health and the National Science Foundation.

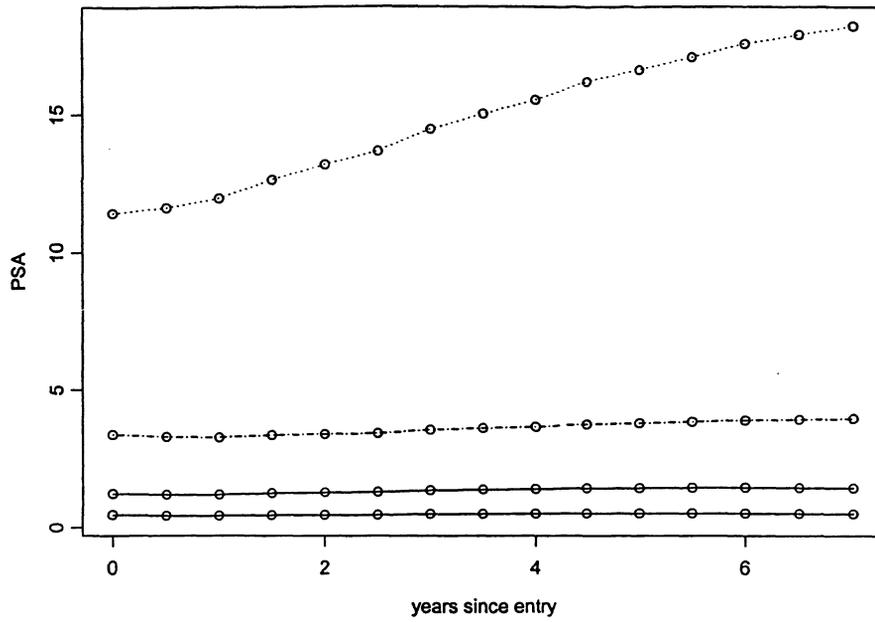
REFERENCES

1. Clark, L. C. et al. 'Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin', *Journal of American Medical Association*, **276**, 1957-1963 (1996).
2. Catalona, W. J., Smith, D.S., Ratliff, T. L. and et al. 'Measurement of prostate-specific antigen in serum as a screening test for prostate cancer', *New England Journal of Medicine*. **324**, 1156-1161 (1991).
3. Carter H. B. et al. 'Estimation of prostatic growth using serial prostate specific antigen measurements in men with and without prostate disease', *Cancer Research*, **52**, 3323-3328 (1992).
4. Emir, B., Wieand, S., Su, J. Q. and Cha, S. 'Analysis of repeated markers used to predict the progression of cancer', *Statistics in Medicine* **17**, 2563-2578 (1998).
5. Agresti, A. *Categorical Data Analysis*, John Wiley & Sons, New York, 1990.
6. Breiman, L., Fridman, J. H., Olshen, R. A. and Stone, C. J. *Classification and Regression Trees*, Wadsworth International Group, Inc., Belmont, 1984.
7. Segal, M. R. 'Tree-structured methods for longitudinal data', *Journal of the American Statistical Association*, **87**, 407-418 (1992).
8. Zhang, H. 'Multivariate adaptive splines for analysis of longitudinal data', *Journal of Computational and Graphical Statistics*, **6**, 74-91 (1997).

9. Laird, N. M. and Ware, J. H. 'Random-effects models for longitudinal data', *Biometrics*, **38**, 963-974 (1982).
10. Belin, T. R. and Rubin, D. B. 'The analysis of repeated-measures data on schizophrenic reaction times using mixture models', *Statistics in Medicine*, **14**, 747-768 (1995).
11. Rubin, D. B. & Wu, Y. 'Modeling schizophrenic behavior using general mixture components', *Biometrics*, **53**, 243-261 (1997).
12. Jacobs, R.A., Peng, F. and Tanner M. A. 'A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures', *Neural Networks*, **10**, 231-241 (1997).
13. Peng, F., Jacobs, R.A. and Tanner M. A. 'Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition', *Journal of the American Statistical Association*, **91**, 953-960 (1996).
14. Qu, Y., Tan, M. and Kutner, M. H. 'Random effects models in latent class analysis for evaluating accuracy of diagnostic tests', *Biometrics*, **52**, 797-810 (1996).
15. Hadgu, A. and Qu, Y. 'A biomedical application of latent class models with random effects', *Applied Statistics*, **47**, 603-616 (1998).
16. Yang, I. and Becker, M. P. 'Latent Variable Modeling of Diagnostic Accuracy', *Biometrics*, **53**, 948-958 (1997)
17. Muthén, B. and Shedden K. 'Finite mixture modeling with mixture outcome using the EM algorithm', *Biometrics*, in press, 1999.
18. McLachlan, G. J. and Krishnan, T. *The EM Algorithm and Extensions*, John Wiley & Sons, New York, 1997.
19. The Math Works, Inc. *MATLAB, Version 5.2.1420*, The Math Works, Inc., Natick, 1998.
20. Catalona, W. J., Smith, D. S. and Ornstein, D. K. 'Prostate cancer incidence in men with serum PSA concentration of 2.6 to 4 ng/ml and benign prostatic examination', *Journal of American Medical Association*, **277**, 1452-1455 (1997).
21. Whittemore, A. S., Lele, C., Friedman, G. D., Stamey, T., Vogelman, J. H. and Orentreich N. 'Prostate-specific antigen as predictor of prostate cancer in black men and white men', *Journal of the National Cancer Institute*, **87**, 354-359 (1995).
22. Pearson, J. D., Morrell, C. H., Landis, P. K., Carter, H. B. and Brant, L. J. 'Mixed-effects regression models for studying the natural history of prostate disease', *Statistics in Medicine*, **13**, 587-601 (1994).

23. Slate, E. H. and Cronin, K. A. 'Changepoint modeling of longitudinal PSA as a biomarker for prostate cancer', in Gatsonis, C., Hodges, J. S., Kass, R. E., McCulloch, R., Rossi, P. and Singpurwalla, N. D. (eds), *Case Studies in Bayesian Statistics III*, Springer-Verlag, New York, 1997, pp444-456.
24. Slate, E. H. and Clark, L. C. 'Using PSA to detect prostate cancer onset: An application of Bayesian retrospective and prospective changepoint identification', in Gatsonis, C., Carlin, B., Carriquiry, A., Gelman, A., Kass, R., Verdinelli, I. and West, M. (eds), *Case Studies in Bayesian Statistics IV*, Springer-Verlag, New York, 1998, pp511-534.
25. Slate, E. H. and Turnbull, B. W. 'Statistical methods for longitudinal biomarkers of disease onset', *Statistics in Medicine*, to appear, 1999.
26. Wu, C. F. J. 'On the convergence properties of the EM algorithm', *Annals of Statistics*, **11**, 95-103 (1983).

Fitted PSA Trajectories for the 4-class Model



Fitted PSA Trajectories for the 3-class Model

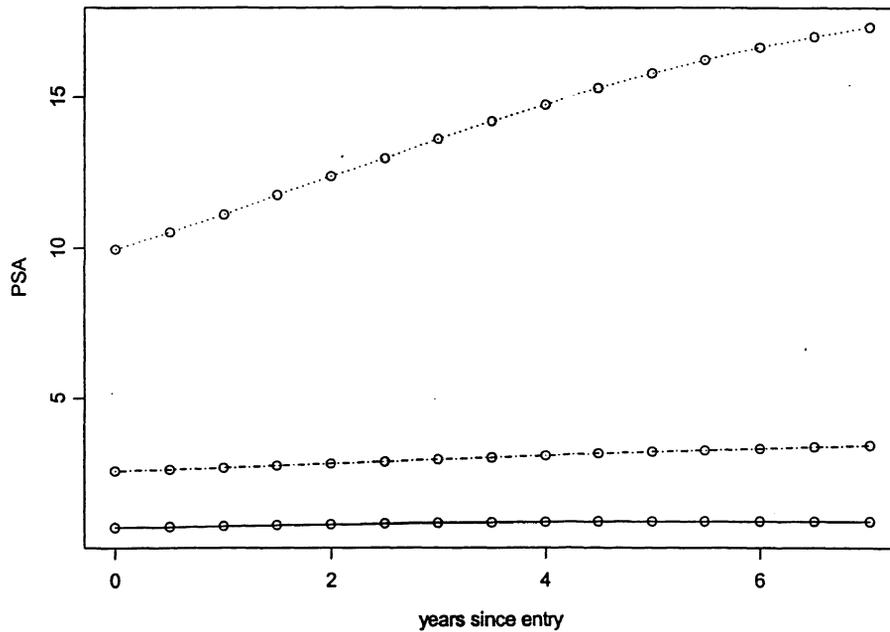
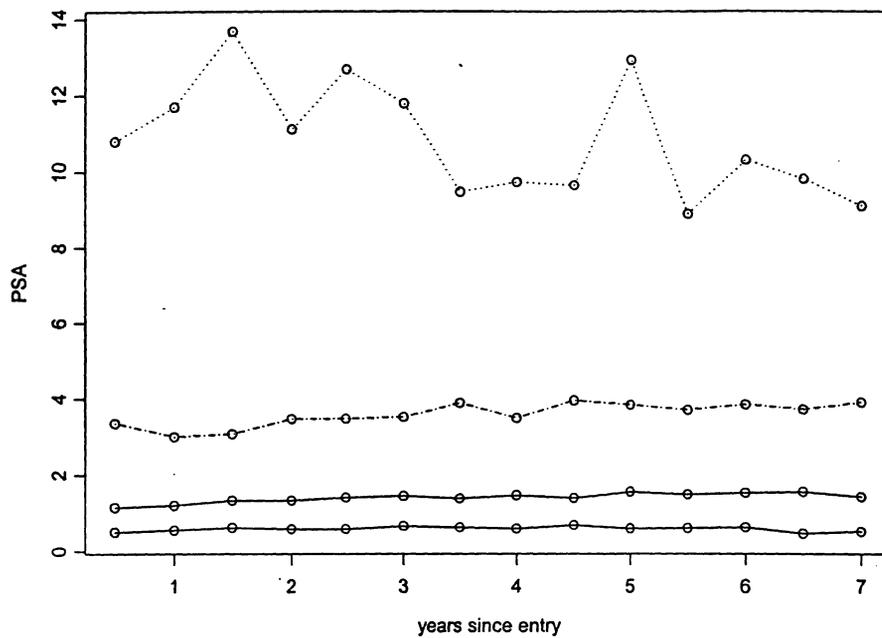


Figure 1: Fitted PSA Trajectories for the 3-Class Solution
solid line: Class "Low", dashed line: Class "Medium", dotted line: Class "High".

Observed Trajectories for the 4-Class Model



Observed Trajectories for the 3-Class Model

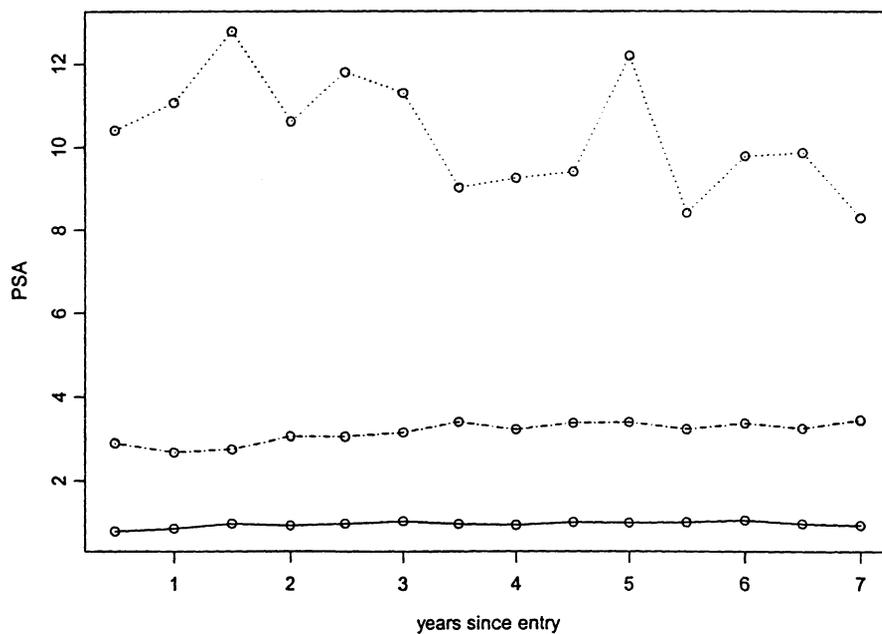


Figure 2: Observed PSA Trajectories for the 3-Class Solution
 solid line: Class "Low", dashed line: Class "Medium", dotted line: Class "High".

residual vs. year

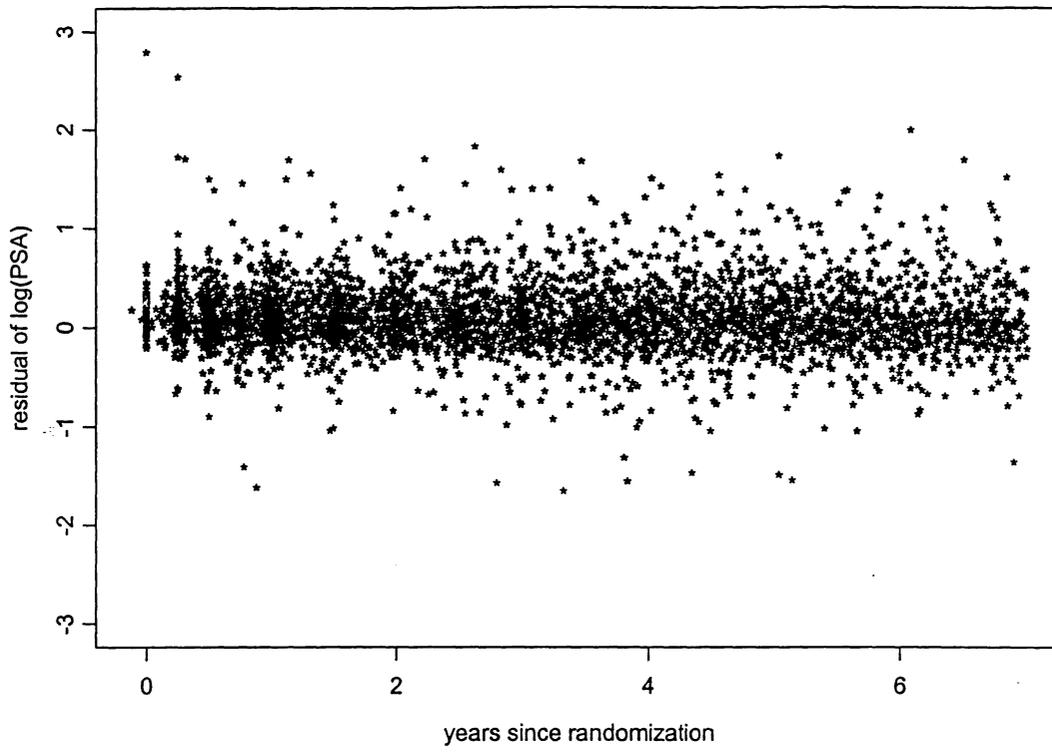


Figure 3: Marginal Residual Plot of PSA