Panel Regression with Unobserved Classes

Mickael Salabasis* Mattias Villani†

SSE/EFI Working Paper Series in Economics and Finance
No. 353
January, 2000

Abstract

We propose a panel regression model with a predetermined and fixed number of classes, where each class is defined by its parameters, but any reference as to which group any observation belongs to is absent. The classes or groups are rationalized by a willingness to attribute some of the observed heterogeneity on a higher level than the individual. The estimation procedures have a distinct Bayesian flavor, relying on the Gibbs sampling scheme for parameter estimation, which has proven effective in situations with missing or latent variables.

*Department of Economic Statistics, Stockholm School of Economics, P.O Box 6501, S-113 83 Stockholm, Sweden, stms@hhs.se. Financial support by the Swedish Council for Research in the Humanities and Social Sciences is gratefully acknowledged.

†Department of Statistics, Stockholm University, S-106 91 Stockholm, Sweden.

1 Introduction

Heterogeneity is a major issue in the econometric analysis of panel data. A quite common practice is to initially address the problem by splitting large panels into smaller parts, conducting the analysis in each subpanel separately. The validity of such a procedure rests on the assumption of any observed heterogeneity not necessarily being individual specific exclusively, but that at least part of it can be attributed to the existence of a, usually small, number of distinct response types. This *behavioral clustering* can be interpreted as a form of heterogeneity, residing at a group level, where the homogenous classes are either actually believed to exist or simply define a convenient structure for the problem at hand.

This common practice discloses a link between econometric analysis of panels and multilevel models, extensively used in social and medical research. Paraphrasing Hox [1995] slightly, one might argue that there should be a place for the multilevel paradigm in the analysis of panel data as economic research often concerns problems that investigate the relationship between agents and the economy, where the general concept is that agents interact with the social and economic contexts to which they are associated. One possible interpretation of this statement is that agents, and specifically their behavior, are influenced by the groups they belong to or associate themselves with, while at the same time said groups may be influenced by their members. The described state of nature can be conceptualized as a hierarchical system, and once established, ignoring groups and group effects may render any standard statistical technique invalid. For the introduction of a systematic statistical approach and development of tools to analyze such structured data, the work of Aitkin and Longford [1986] has played a major role, while more recent advances are covered in Goldstein [1994].

Contrary to multilevel analysis, where the membership, or allocation, of observations to classes is known, the analysis of economic data may be obstructed by the lack of specific information about the hierarchy. Though focus is often set on well defined groups, such as men and women or youths and adults, it is common with more diffuse definition of groups, such as high and low ability or *stayers* and *movers*, where suggestions offered on how to classify observations are frequently vaguely formulated or difficult to apply. Thus, as the definition of groups, and the subsequent classification of individuals as members, may be subtler, we may need to recover them by indirect means. This paper will address the problem of inference in a situation where a class structure is plausible, as such or as an approximation, the number of classes is known and any reference to class membership on the individual level is unavailable.

In section 2, after a brief description of a simple multilevel model, we address the issue of latent class membership. We show how a multi level model may be transformed into a mixture model by assuming a random classification model for group membership. Section 3 focuses on estimation, where the general approach chosen is Bayesian. As the computational burden becomes unmanageable even for small problems the particular approach taken is that of numerical evaluation of the posterior by means of the Gibbs sampler, which neatly solves the implementation of the random allocation model and has proven itself effective in estimation of mixture models. The procedure is then illustrated on a synthetic data set in section 4, and we conclude in section 5 with possible extensions and elaborations.

2 Model

2.1 Multi Level Model

Consider the simple multilevel model

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \varepsilon_{ij},\ i = 1, 2, ..., n;\ j = 1, 2, ..., k,$$

where standard interpretations can be given to the intercepts, slopes and residuals. Subscripts i and j index the level 1 and 2 units, which we will refer to as *individual* and *class* respectively. We may extend the model by including further fixed explanatory variables as well as introducing repeated measurements. Thus for individual i in class j at time t we write

$$y_{ijt} = \mathbf{x}'_{ijt}\boldsymbol{\beta}_j + \varepsilon_{ijt},$$

where the intercept is included in $\boldsymbol{\beta}_j$.

Making the indexing more consistent, we introduce a vector of indicator variables $\mathbf{z}$ such that $z_i = j$ if individual i is a member of class j. With standard assumptions about the distribution of the errors we may write the model for an arbitrary individual i as

$$\mathbf{y}_i \mid \{z_i = j\} \sim N_T\left(\mathbf{X}_i\boldsymbol{\beta}_j, \Sigma_i\right). \tag{1}$$

where $\mathbf{y}_i$ the $(T \times 1)$-vector containing measurements of the response, $\mathbf{X}_i$ the $(T \times p)$-matrix containing measurements of the predictors for individual i, and Σ_i is a possibly non-diagonal covariance matrix which can be restricted to be equal for all individuals within a class or even equal for all individuals in the population.

Not concerning ourselves with more complex covariance structures, as those associated to various random effects specifications, we limit attention to variance components defined as $\Sigma_i = \sigma_j^2\mathbf{I}$ for all members of group j.

2.2 Latent Class Membership

Contrary to multilevel analysis where the hierarchy is known, either as a result of the sampling scheme used or the experimental design, the analysis of economic data may be obstructed by the lack of information about the hierarchy.

When the true membership of observations unknown, that is $\mathbf{z}$ is not observed, conditioning on class is not possible for estimation purposes. A suggested method for circumventing this *missing data* problem is to assume class membership is random and that the probability of belonging to class j is the same for all individuals or, in more formal terms, to assume

$$p(z_i = j) = w_j, \tag{2}$$

which can be interpreted as our sample being a random sample from a mixed population with population proportions given by $w_1, ...w_k$. This simple and realistic description means that the multilevel model in equation (1) may be rewritten as a mixture,

$$\mathbf{y}_i \sim \sum\nolimits_{j=1}^{k} w_j N_T(\mathbf{X}_i \boldsymbol{\beta}_j, \sigma_j^2 \mathbf{I}), \tag{3}$$

reflecting the possibility of any individual potentially being a member of any class.

3 Estimation

3.1 Model Specification

As Bayes estimators for mixtures are well defined, at least in case of proper prior distributions, the general approach in this paper will be Bayesian; for a convincing presentation of the paradigm see Bernardo and Smith [1994].

The starting point of any Bayesian analysis is the joint distribution for all quantities of the model, both known and unknown. Next, conditioning on what has been observed, we obtain the distribution of the unknown quantities conditional on the observed sample, the posterior distribution. The unknown parameters in the proposed model are $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k)$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, ..., \sigma_k^2)$, $\mathbf{w} = (w_1, ..., w_k)$ and the latent $\mathbf{z} = (z_1, ..., z_n)$.

After some reasonable assumptions of independence, the joint distribution of all quantities, known and unknown, can be described by the model

$$p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{w}, \mathbf{z}) = p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z}) p(\boldsymbol{\beta}) p(\boldsymbol{\sigma}^2) p(\mathbf{z}|\mathbf{w}) p(\mathbf{w}). \tag{4}$$

Completing the specification, and in a standard fashion, we select priors

$$\begin{array}{l} p\left(\mathbf{w}\right) \propto w_1^{\delta_1 - 1} \cdots w_k^{\delta_k - 1}, \\ p\left(\boldsymbol{\beta}_j\right) \propto N_p\left(\boldsymbol{\mu}_0, \sigma_0^2 \mathbf{I}\right), \\ p\left(\sigma_j^{-2}\right) \propto Ga(a, b), \end{array} \tag{5}$$

that is an informative Dirichlet prior for the population weights and proper conjugate priors for the regression parameters. The choice of a conjugated model is again one of convenience, any proper prior will do fine. The selection of parameters $(\boldsymbol{\delta}, \boldsymbol{\mu}_0, \sigma_0^2, a, b)$ is either based on prior information available or, lacking any relevant information, in such a way as to make the priors appropriately vague.

A potential, and often real, problem with specification (4) is that, lacking any prior information, the model is invariant to permutations of the class labels, and is therefore unidentified. In the literature, this phenomenon is called *label switching* and solutions offered include the introduction of some identifiability constraints, various reparametrizations of the model, and post processing the output, of which the first is often preferred and chosen here. A natural choice of parameter to impose an identifiability constraint on is $\boldsymbol{\beta}$. One such restriction could be

$$\beta_1^{(l)} < \beta_2^{(l)} < \ldots < \beta_k^{(l)},$$

where $\beta_j^{(l)}$ is the *l-th* element of $\boldsymbol{\beta}_j$. Thus, groups are defined in ascending order of the *l-th* element, though $\mathbf{w}$ or Σ, in case of group specific variances, could also be used.

In accordance with the above the model is rewritten as

$$p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{w}, \mathbf{z}) = p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z}) p(\boldsymbol{\beta}) p(\boldsymbol{\sigma}^2) p(\mathbf{z}|\mathbf{w}) p(\mathbf{w}) I_{\beta_1^{(l)} < \beta_2^{(l)} < \ldots < \beta_k^{(l)}}, \tag{6}$$

where $I_{\beta_1^{(l)} < \beta_2^{(l)} < \ldots < \beta_k^{(l)}}$ is an indicator function for the restricted $\boldsymbol{\beta}$-space.

Though the conditioning of equation (6) on the sample, for particular prior distributions, results in a posterior distribution that can be written in closed form, the computing time needed becomes prohibitively large even for a small number of groups k and moderate sample size n as it involves the evaluation of k^n terms.

With the direct Bayesian approach disqualified, the approach taken is that of numerical evaluation of the posterior by the Gibbs sampler, for details and further references see Gilks, Richardson and Spiegelhalter [1996] and especially Robert [1996] in that volume.

3.2 Full Conditional Posteriors

Implementation of the Gibbs sampler requires calculation of the full conditional posteriors of all parameters of interest.

3.2.1 Full conditional posterior of $\boldsymbol{\beta}_j$

For class j we have the following full conditional posterior

$$\boldsymbol{\beta}_j|\mathbf{y}, \boldsymbol{\beta}_{-j}, \boldsymbol{\sigma}^2, \mathbf{w}, \mathbf{z} \sim N_p(\bar{\boldsymbol{\beta}}_j, \Omega_j),$$

where

$$\begin{aligned} \Omega_j &= [\sigma_j^{-2}\tilde{\mathbf{X}}_j'\tilde{\mathbf{X}}_j + \sigma_0^{-2}\mathbf{I}]^{-1} \\ \bar{\boldsymbol{\beta}}_j &= \Omega_j[\sigma_j^{-2}\tilde{\mathbf{X}}_j'\tilde{\mathbf{y}}_j + \sigma_0^{-2}\boldsymbol{\mu}_0], \end{aligned}$$

where $\tilde{\mathbf{y}}_j$ is the n_jT-dimensional vector containing the stacked $\mathbf{y}_i$ for the n_j observations currently allocated to class j and $\tilde{\mathbf{X}}_j$ contains their respective $\mathbf{X}_i$ stacked, and $\boldsymbol{\beta}_{-j}$ the collection of all parameter vectors with the exception of $\boldsymbol{\beta}_j$.

3.2.2 Full conditional posterior of σ_j^{-2}

For the variances we get the full conditional posterior

$$\sigma_j^{-2}|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2_{-j}, \mathbf{w}, \mathbf{z} \sim Ga(a + \frac{n_jT}{2}, b + \frac{1}{2}(\tilde{\mathbf{y}}_j - \tilde{\mathbf{X}}_j\boldsymbol{\beta}_j)'(\tilde{\mathbf{y}}_j - \tilde{\mathbf{X}}_j\boldsymbol{\beta}_j))$$

where n_j, $\tilde{\mathbf{y}}_j$ and $\tilde{\mathbf{X}}_j$ are defined as above. For a common variance specification suppress the index j.

3.2.3 Full conditional posterior of w

For the population weights the full conditional posterior is given by

$$\mathbf{w}|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{z} \sim Dirichlet(n_1 + \delta_1, ..., n_k + \delta_k),$$

where n_j defined as above and $\delta_1, ..., \delta_k$ are hyperparameters reflecting a belief about the relative sizes of the groups, as well as the strength of that belief.

3.2.4 Full conditional posterior of z

For the allocation we get

$$\pi(z_i = j|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{w}, \mathbf{z}_{-i}) \propto w_j N_T(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j, \sigma_j^2\mathbf{I}).$$

The allocation of observations to groups is carried out using the discrete distribution implied by $\{\pi(z_i = j|\cdot), j = 1, ..., k\}$, for details see Press [1989].

3.3 Gibbs Sampler Implementation

With the missing data structure introduced as in section 2.2 and the full conditional posteriors derived in section 3.2 the implementation of the Gibbs sampler is straightforward.

Given some initial allocation and starting values for all model and prior parameters, the Gibbs sampler proceeds as follows

Step 1. Sample new weights $\mathbf{w}$ based on the current cardinality of each class and the relevant posterior in section 3.2.3.

Step2. Cycle through all n observations, calculate for each observations k numbers in accordance with section 3.2.4 and sample a new allocation z_i from the resulting individual discrete distribution. Recalculate current cardinality n_j of each class.

Step 3. Given the allocation from step 2 create $\tilde{\mathbf{y}}_j$ and $\tilde{\mathbf{X}}_j$, the stacked vectors and matrices. Cycle through all k classes and sample new $\boldsymbol{\beta}_j$ from the posterior distribution in section 3.2.1 such that the order condition in model (6) is not violated.

Step 4. Given the allocation from step 2 create the residuals $\tilde{\mathbf{y}}_j - \tilde{\mathbf{X}}_j$. Cycle through all k classes and sample new σ_j^{-2} from the posterior distribution in section 3.2.2.

Repeat until convergence of the Gibbs sampler.

4 Illustration

To illustrate the suggested procedure we consider a synthetic data set supporting the model without being trivial, and designed in such a way that we may gain some insight into problems that may be encountered in practice.

4.1 Simple Model Specification

We generate a sample of $n = 200$ individuals and $T = 5$ measurements from a simple model with $k = 2$ classes, $p = 2$ regressors, an intercept and a linear trend, common variance and parameters in $\boldsymbol{\beta}$ selected in such a way that the

generated data will support the model without being trivial.

$$\begin{aligned} N &= 200, T = 5, p = 2, k = 2 \\ w_1 &= P(z_i = 1) = 0.40 \\ \mathbf{y}_{it} &\mid \{z_i = j\} = \mathbf{X}\boldsymbol{\beta}_j + \varepsilon_{it} \\ \mathbf{X} &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}' \\ \boldsymbol{\beta}_1 &= (1.00, 1.00)', \boldsymbol{\beta}_2 = (1.25, 0.75)' \\ \varepsilon_i &\sim N_5\left(0, \sigma^2 \mathbf{I}_5\right), \sigma^2 = 1. \end{aligned} \tag{7}$$

A typical sample is illustrated in Figure 1, where each point represents the individual estimates of intercept and slope. As expected the variation along the intercept dimension is greater, but the two groups are discernible, though not accentuated.

<Figure 1 about here>

4.2 Some Practical Considerations

To apply the Gibbs sampler we need to complete the prior specifications with parameter values δ_1 and δ_2 for the weights, $\boldsymbol{\mu}_0$ and σ_0^2 for the regression parameters, a and b for the error variance, select some starting values, decide on burn-in length and stopping time, as well as choose a regressor for the identifying restriction.

Reflecting a prior belief of equal population proportions we select $\delta_1 = \delta_2 = 10$, for the regression parameters we set $\boldsymbol{\mu}_0 = [0, 0]'$ and $\sigma_0^2 = 100$, and for the variance, or rather the precision, $a = b = 0.01$. As there is only one explanatory variable, the trend, it is selected for identification purposes. Thus we label groups in ascending order of their trend coefficient, and restrict any sampled values so that $\boldsymbol{\beta}_1^{(1)} < \boldsymbol{\beta}_2^{(1)}$.

In theory starting values are of minor importance as they will not affect the stationary distribution of an irreducible chain. In practice, and especially for slow mixing chains, some care should be exercised, to avoid lengthy burn-in times, for some suggestions see Gelman and Rubin [1992]. Thus, to obtain starting values we begin by estimating $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ unconditionally, setting

$$\begin{aligned} \boldsymbol{\beta}_{1,(0)} &= \hat{\boldsymbol{\beta}} - \kappa \cdot diag\left(X'X\right)^{-1} \\ \boldsymbol{\beta}_{2,(0)} &= \hat{\boldsymbol{\beta}} + \kappa \cdot diag\left(X'X\right)^{-1} \\ \sigma^2_{(0)} &= \hat{\sigma}^{-2} \\ w_1 &= w_2, \end{aligned}$$

where κ is set to 0.05, but any small value will do. An initial allocation is chosen based on the relative distances of the individual estimates of intercept and slope from $\boldsymbol{\beta}_{1,(0)}$ and $\boldsymbol{\beta}_{2,(0)}$.

The length of necessary burn-in depends on starting values, the rate of convergence of the Gibbs sampler to the desired stationary distribution and the degree of approximation demanded. In theory it may be determined analytically though, in most cases, the necessary calculations present an insurmountable problem. In practice visual inspection of output is the most commonly used method while other available rules of thumb include the suggestion by Geyer [1992] that, with reasonable starting values, a burn-in length of between one and two percent of total length is adequate.

We also need to determine a stopping time, the goal being adequate final precision. For a small problem as above, where every iteration is very fast, the issue is of minor importance, but with increasing sample size, number of measurements, classes and covariates, the issue of computational efficiency becomes real. For some guidance on formal methods of deciding stopping time we refer to Raftery and Lewis [1996]. As economy is not really an issue for the example at hand, and adhering to Geyer's rule of thumb, we discard the first 1000 iterations as burn-in, and run the Gibbs sampler for another $100,000$ iterations.

4.3 Results and Inference

Any statistic based on the Gibbs sampler output is valid for inference as long as the chain has converged. Addressing the question of convergence we may inspect the evolution of the posterior mean of parameters of interest. In Figure 2, the posterior means of $\boldsymbol{\beta}$ are plotted, and we note a relatively rapid convergence, within the first 20000 iterations for all coefficients. The same plots for the weights and variance display similar behavior. The convergence properties of the parameter estimates were further corroborated by their estimated Monte Carlo variances, evaluated every 20000 iterations by running a large number of parallel chains.

<Figure 2 about here>

The posterior distributions of the parameters of interest, as in Figure 3 for the regression coefficients, are a concise and complete way to summarize the results of the Gibbs sampler. Visual inspection of these particular posterior distributions seem to indicate a very slight truncation for the slopes, which is probably a result of the ordering condition. A more complete and manageable presentation of the results is presented in Table 1, with some selected sample percentiles, means and standard deviations for the posterior distributions of $\boldsymbol{\beta}$, w_1, and σ^2. To facilitate the assessment of the results we also report

the model and sample equivalents, estimated assuming known allocation. As expected, the Gibbs sampler managed to discriminate the two classes, resulting in good estimates overall.

<Figure 3 about here>

<Table 1 about here>

A variable with interesting inferential possibilities is the introduced latent allocation $\mathbf{z}$. One possibility is to use the posterior modal allocation and other relevant posterior distributions to fit any preferred classification model, to be used for quick membership predictions of new individuals. Another possibility, equally valid, would be to actually fit any such classification model within the Gibbs sampler. This is illustrated in Figure 4, where three labeled contours for the probability of being a member of the first class, estimated in each iteration by a linear discriminant which relies on calculations similar to those performed for the allocation in section 3.2.4, have been averaged. We also compare the true allocation with the posterior allocation of observations to class one. This comparison is not possible in applications but is done to illustrate the degree of correspondence between actual and sampled allocation.

<Figure 4 about here>

5 Remarks

5.1 Problems and Pitfalls

Several things influence the performance of the proposed model and estimation procedure, where most problems are related to the definition and identification of classes.

Critical factors for a good performance is the similarity of members within a class and the distinctiveness of classes relative each other. For the model considered, the similarity of members within a class is primarily governed by the precision of the individual estimates of the regression coefficients, where, all else equal, smaller sample variation in individual estimates on average implies clearer definition of classes. Still, within class similarity is a relative concept as it depends in a sense on the distinctiveness of classes, as measured by some location distance. The model may perform very well even when the within class variation is large if the classes are located far from each other. This is illustrated in Figure 5 for the model used in section 4, where confidence ellipsoids for the two classes are plotted. Holding the means constant, decreasing similarity within classes, due to for instance a larger error variance or fewer observations per individual, would imply wider confidence regions,

a greater overlap between classes, and difficulties in estimation due to inadequate distinctiveness. Thus performance is obviously problem dependent and it may be difficult to state prior to estimation what the risks are.

<Figure 5 about here>

With a fix number of classes k, it is necessary that data actually supports the model. By this we mean not only that k is correctly specified but also that the relation of within and between variation, as defined above, is such that we may have confidence in the Gibbs sampler actually picking out the classes. Otherwise, when none or too few observations are allocated to any class we will have problems sampling new values and get stuck in areas of low probability; this despite of the theoretical irreducibility of the chain. Still this computational problem may occur even under the best of circumstances.

Another cause of concern is identification and identifying restrictions. As mentioned in section 3.1, when estimating mixtures with the Gibbs sampler, if not properly identified we may have trouble discriminating between the $k!$ possible different posterior modes. When using an order constraint, as we do, there may be better or worse choices of element to order upon. Again the optimal decision depends on both the within and between variation, as well as the actual distance between class loci, and it may be difficult to form an opinion *a priori*.

The allocation step in the Gibbs sampler, and in particular the calculation of the individual discrete distributions of membership, is also a source of potential problems of numerical nature. Coupled with the trapping state complication mentioned above, outliers may have a severe detrimental effect on performance.

Otherwise, some limited experimentation indicates, as expected, that the performance is enhanced with increasing sample size, increased number of measurements, and distinctiveness of mixture components.

5.2 Extensions and Elaborations

Improved performance and increased richness of specification, are two dimensions along which extensions and elaborations may be discussed.

As discussed in section 5.1, we may experience practical and technical problems when estimating the mixture. Practical problems are associated primarily with the definition of classes while the technical problems are of computational nature, frequently connected to the former. A source to which many of the problems can be traced is class membership and the random membership model presented in section 2.2. The mechanism at work is simple, and may result in the Gibbs sampler being trapped in areas of low probability while exploring them, as is necessary. In the literature, the main

direction suggested for solving both problems are various reparametrizations as in Gilks and Roberts [1996] and Robert [1996]. Another suggested direction, perhaps more natural and certainly offering very interesting possibilities, is through the introduction of the number of classes into the analysis, as in Green [1995] and Richardson and Green [1997]. A third possibility, when auxiliary information is available, would be to extend the membership model from the random used to something more elaborate like the probit, where the work of Chib and Greenberg [1998] might offer valuable insights.

An obvious venue for increased model flexibility is the specification of the covariance matrix, which could be generalized to handle more intricate structures and extended to specifications taking into account the time series aspect of panels. The obvious way to do this is through the prior specification which may be altered to comply with one or two way random effects, at a moderate computational cost.

Finally, the idea of the class structure model, presented in section 3, is that the effect of the regressors in $\mathbf{X}$ operate at a group level, the effect of the variable depending on group membership. The flexibility of the model can be further increased by the introduction of individual and population effects, that is independent variables operating at an individual and population level. The model for an individual could then be written as

$$\mathbf{y}_i \mid \{z_i = j\} \sim N_T(\boldsymbol{\alpha}_i \mathbf{X}_{1i} + \boldsymbol{\beta}_j \mathbf{X}_{2i} + \boldsymbol{\gamma} \mathbf{X}_{3i}, \Sigma_i), \tag{8}$$

where variables classified as $\mathbf{X}_1$ are interpreted as random or fixed individual effects, while variables classified as $\mathbf{X}_3$ would correspond to the independent variables of a standard panel regression, that is variables with an effect that does not depend on group membership and is the same for all individuals. What makes model (8) interesting is the fact that it can be used to specify anything from a simple panel regression model, where all independent variables are classified as being of type 3, to something resembling a random coefficient model, where all independent variables are classified as being of type 1, with an option of mixing any or both with variables of type 2 to create hybrid models of varying degrees of complexity.

References

Aitkin, M. and Longford, N.: 1986, Statistical modelling in school effectiveness studies., *Journal of the Royal Statistical Association* **149**, 1–43.

Bernardo, J. M. and Smith, A. F. M.: 1994, *Bayesian Theory*, Wiley.

Chib, S. and Greenberg, E.: 1998, Analysis of multivariate probit models, *Biometrika* **85**(2), 347–361.

Gelman, A. and Rubin, D. R.: 1992, Inference from iterative simulation using multiple sequences, *Statistical Science* **7**, 457–511.

Geyer, C. J.: 1992, Practical markov chain monte carlo, *Statistical Science* **7**, 473–511.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J.: 1996, Introducing markov chain monte carlo, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, pp. 1–19.

Gilks, W. R. and Roberts, G. O.: 1996, Startegies for improving MCMC, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, pp. 89–114.

Goldstein, H.: 1994, *Multilevel Statistical Models*, E Arnold, New York.

Green, P. J.: 1995, Reversible jump markov chain monte carlo computation and bayesian model determination, *Biometrika* **82**, 711–732.

Hox, J. J.: 1995, *Applied Multilevel Analysis*, TT-Publikaties, Amsterdam.

Press, S. J.: 1989, *Bayesian Statistics: Principles, Models, and Applications*, Wiley, New York.

Raftery, A. E. and Lewis, S.: 1996, Implementing MCMC, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, pp. 115–130.

Richardson, S. and Green, P. J.: 1997, On bayesian analysis of mixtures with an unknown number of components, *Journal of the Royal Statistical Society* **59**(4), 731–792.

Robert, C. P.: 1996, Mixtures of distributions: Inference and estimation, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, pp. 441–463.

Table 1: Summary of Gibbs sampler results; selected percentiles, mean and standard deviation for regression coefficients, variance and weight. Model and sample values, where the latter were estimated with OLS assuming knowledge of the true model and individual class membership, are shown to facilitate comparison and appraisal.

	Posterior percentiles, mean and standard deviation								
	$m_{0.05}$	$m_{0.25}$	$m_{0.50}$	$m_{0.75}$	$m_{0.95}$	mean	std	true	sample
$\boldsymbol{\beta}_1^{(1)}$	0.875	0.950	0.998	1.041	1.097	0.993	0.068	1.000	0.992
$\boldsymbol{\beta}_1^{(2)}$	0.690	0.856	0.977	1.105	1.308	0.985	0.189	1.000	1.006
$\boldsymbol{\beta}_2^{(1)}$	1.164	1.202	1.229	1.258	1.304	1.231	0.042	1.250	1.238
$\boldsymbol{\beta}_2^{(2)}$	0.562	0.702	0.792	0.880	1.011	0.789	0.137	0.750	0.767
w_1	0.256	0.338	0.403	0.475	0.586	0.410	0.100	0.400	0.405
σ^2	0.960	1.006	1.041	1.077	1.133	1.043	0.053	1.000	1.039

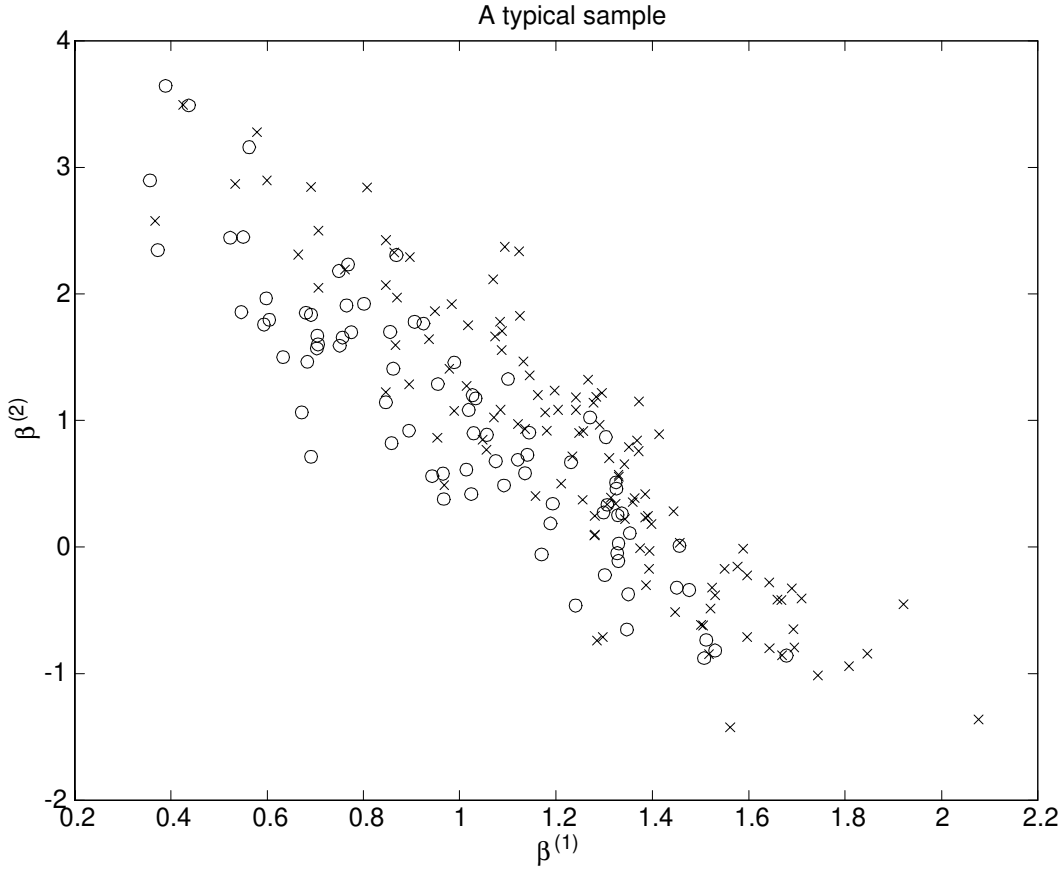

Figure 1: A typical sample of $n = 200$ individuals from the simple model described in (7). Points represent individual estimates of the regression coefficients.
o : observations actually sampled from class 1
x : observations actually sampled from class 2

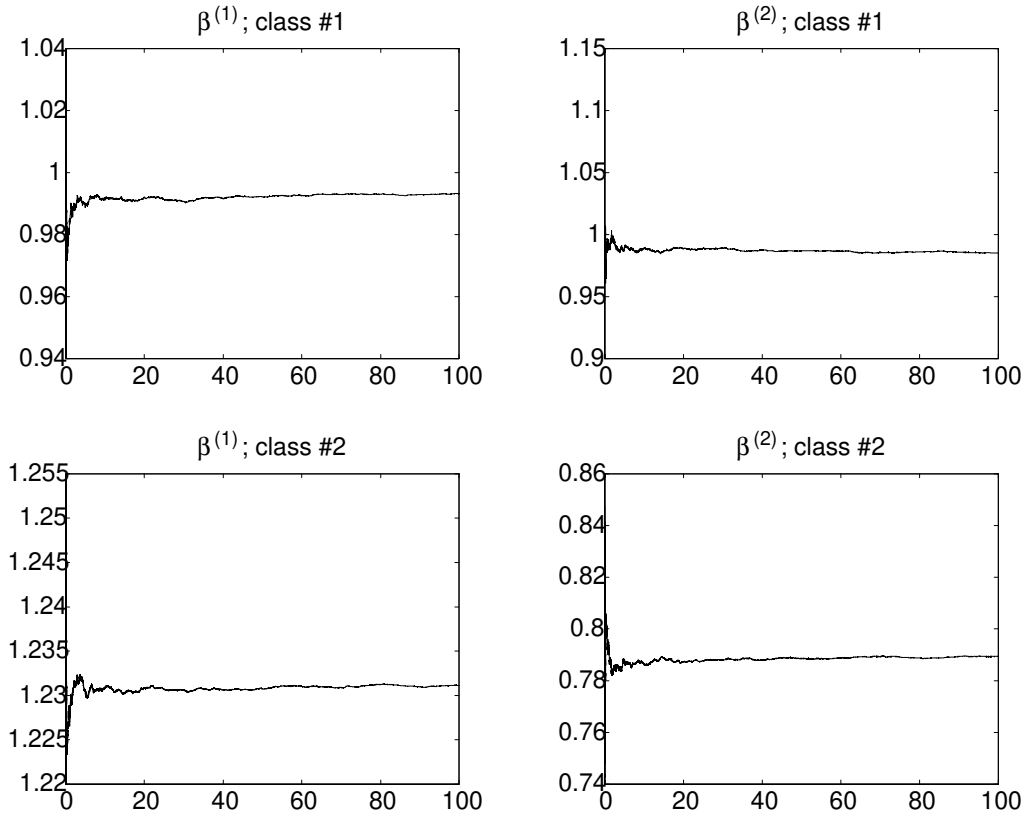

Figure 2: Posterior mean evolution of regression coefficients, iterations in thousands.

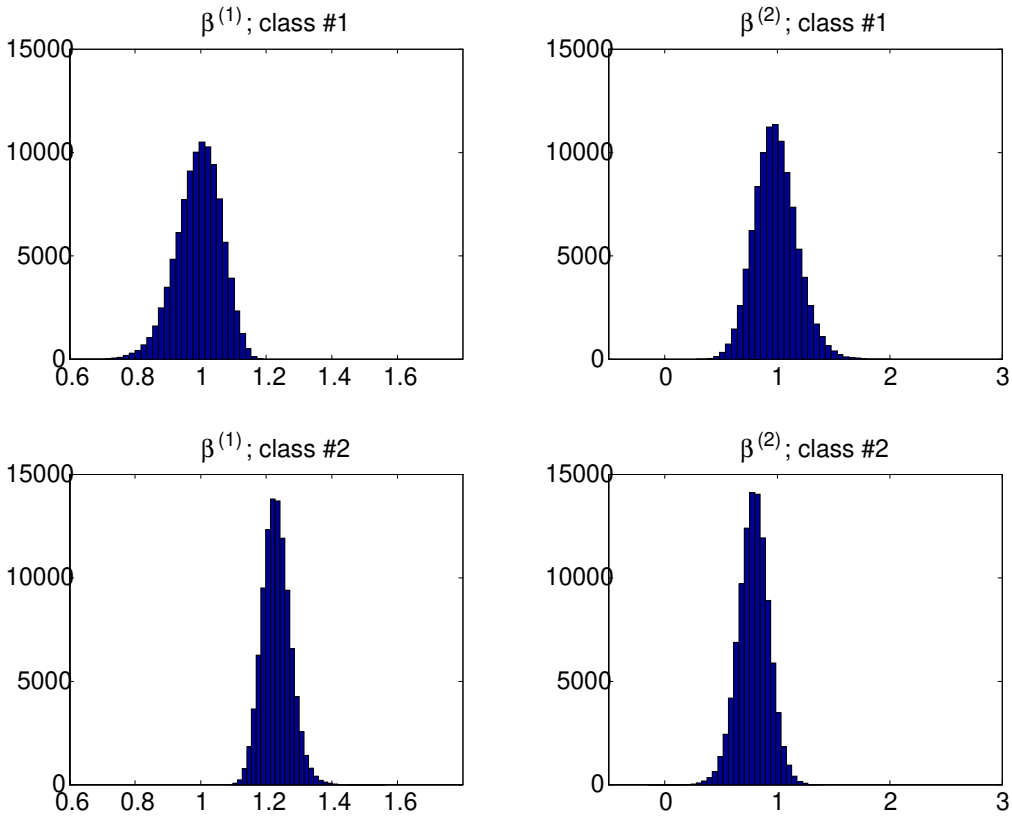

Figure 3: Summary of GIBBS sampler output for all regression coefficients; both groups. The slight truncation in the distributions for the slopes is an effect of the order restriction.

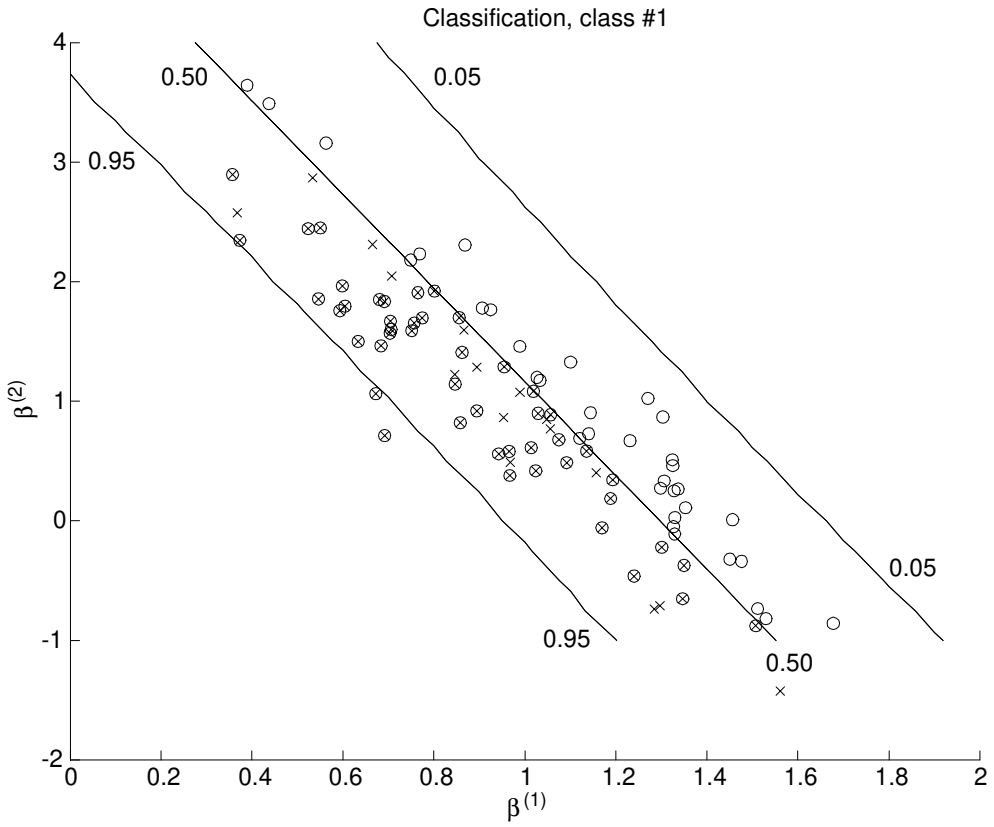

Figure 4: Classification, true and posterior for the first class, with three contours for the probability of an individual being a member of the first class equal to $0.05, 0.50$ and 0.95. The contours are based on a linear discriminant, calculated in each Gibbs sampler iteration for sampled values on all relevant quantities, and averaged over the whole cycle.
o : observations actually sampled from class 1
x : observations with posterior allocation to class 1

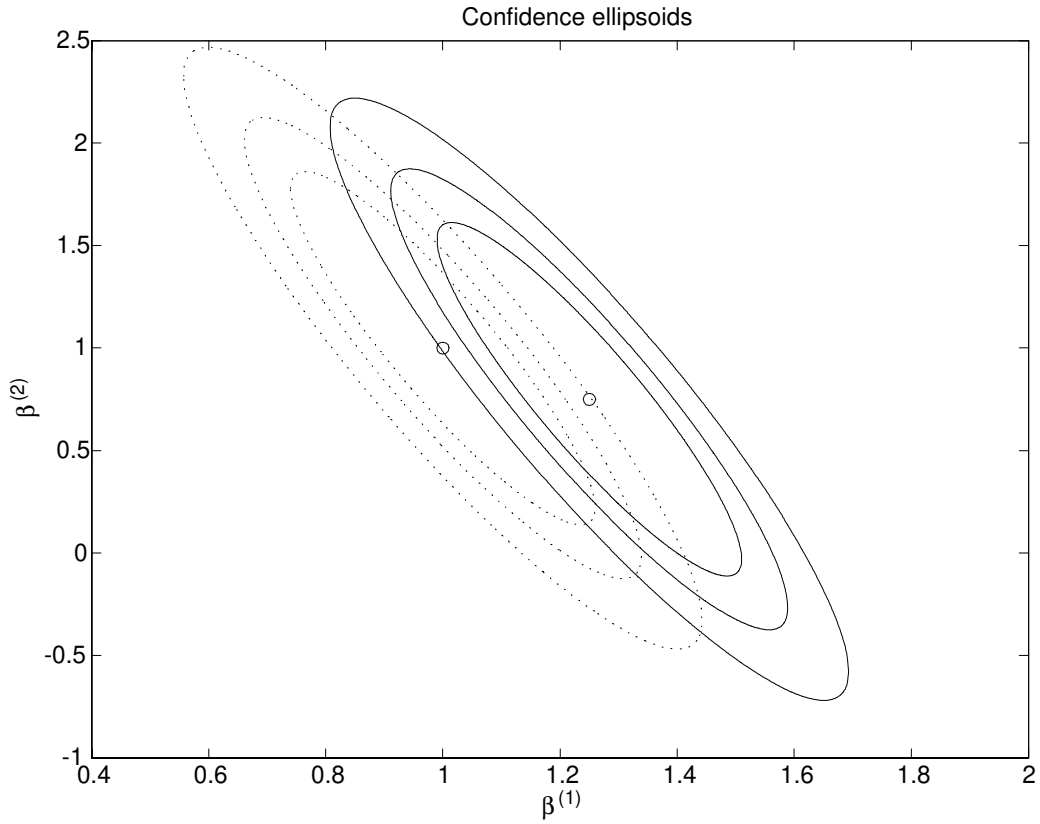

Figure 5: Confidence ellipsoids for both classes, dotted lines for class 1 and solid lines for class 2, where both class means are marked with a circle. The probability of a sampled individuals' estimated $\boldsymbol{\beta}$ falling inside its relevant outer, middle and inner ellipse is 0.90, 0.75, and 0.50 respectively.