Bayesian Factor Models in Characterizing Molecular Adaptation

Saheli Datta
saheli@ams.ucsc.edu

Raquel Prado
raquel@ams.ucsc.edu

Abel Rodríguez
abel@ams.ucsc.edu

Abstract

Assessing the selective influence of amino acid properties is important in understanding evolution at the molecular level. A collection of methods and models have been developed in recent years to determine if amino acid sites in a given DNA sequence alignment display substitutions that are altering or conserving a prespecified set of amino acid properties. Residues showing an elevated number of substitutions that favorably alter a physicochemical property are considered targets of positive natural selection. Such approaches usually perform independent analyses for each amino acid property under consideration, without taking into account the fact that some of the properties may be highly correlated. We propose a Bayesian hierarchical regression model with latent factor structure that allows us to determine which sites display substitutions that conserve or radically change a set of amino acid properties, while accounting for the correlation structure that may be present across such properties. We illustrate our approach by analyzing simulated data sets and an alignment of lysin sperm DNA.

1 Introduction

Several methods for detecting departures from neutrality at the molecular level are based on studying patterns of polymorphism in protein coding genes. Specifically, methods such as those in Li [1993], Suzuki and Gojobori [1999], Nei and Kumar [2000], Yang et al. [2000a], Yang and Nielsen [2002], Yang and Swanson [2002], Ronquist and Huelsenbeck [2003], Huelsenbeck and Dyer [2004], Suzuki [2004] and Anisimova and Kosiol [2009], among others, compare nonsynonymous (amino acid replacing) to synonymous (amino acid conserving) mutation rates and conclude that a given DNA region, residue, or branch along an evolutionary tree (phylogeny) is a target of positive natural selection if it displays an excess of nonsynonymous over synonymous mutations.

Methods that take into account changes in key physicochemical amino acid properties induced by the nonsynonymous mutations have also been proposed and used to assess positive selection at the molecular level. For instance, Xia and Li [1998], McClellan and McCracken [2001], and McClellan et al. [2005] use calculations of expected random distributions, under the assumption of neutrality, of possible amino acid changes based on fixed differences between residues given a particular amino acid property. Alternative approaches based on stochastic models of sequence evolution are developed in Sainudiin et al. [2005] and Wong et al. [2006]. These approaches model the evolution of protein coding sequences as continuous-time Markov chains with rate matrices that distinguish between mutations that alter a given amino acid property and those that conserve it. More specifically, Sainudiin et al. [2005] divides the mutations into two groups: a class of property-conserving substitutions — which includes synonymous and nonsynonymous substitutions that do not alter the amino acid property — and a class of property-altering substitutions — which are always nonsynonymous. Wong et al. [2006] then generalizes the idea to allow for three types of mutations: synonymous, property-conserving nonsynonymous, and property-altering nonsynonymous. It should be noted that both these methods depend on user-specified partitions of the amino acids for one or more properties, and so the results are not robust to the choice of partitions.

Recently, Datta et al. [2010] considered a Bayesian hierarchical regression model that compares the observed amino acid distances to the expected distances under neutrality for a given set of amino acid properties and incorporates mixture priors for variable selection. Unlike the approaches of Sainudiin et al. [2005] and Wong et al. [2006], which are based on partitions that categorize differences in amino acid changes, the model of Datta et al. [2010] directly describes the absolute distances under the different physicochemical properties. Directly modeling the distances, instead of amino acid partitions based on them, has the advantage of avoiding biases that may arise when the values of the properties are close to the class limits. For example, amino acids could be partitioned according to their Hydrophobicity index into those with indexes below 1.0 and those with indexes above 1.0. Under such a partition, Gln (1.07) and Tyr (2.65) would belong to the same class while Gly (0.47) would be in a different class, however, the absolute difference between Gln and Gly is smaller than that between Gln and Tyr. In addition, by incorporating hierarchical mixture priors for variable selection, the model in Datta et al. [2010] is capable of identifying neutral, conserved and positively selected sites while automatically adjusting for multiple comparisons and borrowing information across properties and sites.

The approaches listed above differ considerably with respect to their modeling assumptions and their capabilities in terms of whether they are able to properly adjust for multiple comparisons in cases when molecular adaptation needs to be assessed at several amino acid sites. However, a

common feature of all such methods is that they implicitly assume that properties are independent from each other in terms of their effect on evolution. This is typically an unrealistic assumption unless the properties included in the analysis are carefully chosen. Indeed, a review of the amino acid index database (available at `http://www.genome.jp/dbget/aaindex.html`), which lists more than 500 amino acid properties, shows that a large number of them are highly correlated. Therefore, the finding that a specific property influences evolution at a given site provides information about the effect that other properties may have on the same site, a feature that can be exploited to improve the ability of the model to detect conserved, neutral and positively selected sites. With this insight in mind, we extend the Bayesian hierarchical regression model of Datta et al. [2010] by adding a latent factor structure to jointly model a large number of amino acid properties, several of which may be correlated. The idea of using factor models and principal component analysis to describe the physical properties of the amino acids has been used in the past (e.g., Kidera et al. [1985], Atchley et al. [2005]); however, our approach differs from this previous work in that we do not try to explain correlation in general, but to exploit those correlations to improve detection of those sites subject to natural selection. In fact, the correlations we observe in our models can be very different from those computed from the raw amino acid scores, as they are affected by factors such as codon usage bias. In addition to handling several amino acid properties via the latent factor structure, the model we propose allows us to determine if such properties are being conserved, neutral or radically changed at the amino acid level while properly adjusting for multiple comparisons when several amino acid sites are considered.

The paper is organized as follows. Section 2 describes the procedure to compute distances based on amino acid properties, as well as the structure of the model. Section 3 briefly summarizes some features of the Markov chain Monte Carlo (MCMC) algorithm used to fit the model. Section 4 presents a series of simulation studies that illustrate the performance of the model and how the inference of the latent factor structure is affected by the actual number of strongly conserved and neutral sites. Finally, Section 5 applies the model to the well studied DNA alignment of Lysin protein from abalones and Section 6 gives concluding remarks.

2 Model specification

2.1 Amino acid distances

The structural and functional role of an amino acid in a gene determines its ability to freely change. For example, suppose that certain sites in a protein require a hydrophobic amino acid to maintain its normal function. At such sites, more frequent substitutions occur between amino acids that

have similar values on the hydrophobic scale Xia and Li [1998]. This suggests that we can understand the significance of a given property on molecular evolution by comparing the magnitude of the change in such property under the assumption that a site is neutral under selection (which we call the *expected* change) with the magnitude of the change actually observed in the data (the *observed* change). The reasoning behind this is that a replacement between two amino acids with similar physicochemical characteristics may not change the phenotype in the same way that a replacement between two amino acids that are radically different does (Hughes et al. [1990], Zhang [2000]).

More specifically, our data consist of *observed* and *expected* amino acid distances derived from a DNA sequence alignment, a specific phylogeny, a stochastic model of sequence evolution, and a predetermined set of physicochemical amino acid properties. The distances are obtained by inferring the ancestral sequences from a DNA sequence alignment assuming a fixed phylogeny and a given model of sequence evolution. The main focus of this paper is the development and implementation of models that allow us to make inferences on the latent effects that several amino acid properties may have on natural selection at the molecular level for given a phylogeny and a stochastic model of sequence evolution. Therefore, we do not consider uncertainty in the alignment, phylogeny, or at the ancestral sequence level. For analyses that take into account these uncertainties and further discussion on this issue see Datta et al. [2010].

Our proposed hierarchical model compares the observed amino acid distances, inferred from ancestral sequences based on a given phylogeny, to the expected distances, computed under a process consistent with neutrality, for a given set of amino acid properties. Both the expected and the observed distances are calculated for each amino acid property and for each site showing nonsynonymous substitutions. In order to calculate the observed distances, we first infer the ancestral sequences under a specific substitution model and a given phylogeny. Nonsynonymous substitutions are then counted by comparing DNA sequences between two neighboring nodes in the phylogeny. The observed mean distance, $y_{i,j} \equiv D_O^{i,j}$ for site i and property j is obtained as the mean absolute difference in the property scores due to all nonsynonymous substitutions at site i. Only those sites with at least one nonsynonymous change at the ancestral level are retained for further analysis.

To compute the expected distances, note that each codon can mutate to one of at most nine alternative codons through a single nucleotide substitution [Xia and Li, 1998], of which only some mutations are nonsynonymous (changes to stop codons are ignored). Denote by N_k, the number of nonsynonymous mutations possible through a single nucleotide change, corresponding to a particular codon k ($k = 1, \ldots, 61$). Let $D_{k,l}^{i,j}$ denote the absolute difference in property j between nonsynonymous codon pairs at

site i differing at one codon position, where $l = 1, \ldots, N_k$. The frequency of codon k at a particular site i in the DNA sequence under study is denoted by F_k^i. Then, the expected mean distance for a particular site i and a given property j is given by

$$x_{i,j} \equiv D_E^{i,j} = \frac{\sum_{k=1}^{61} F_k^i \sum_{l=1}^{N_k} D_{k,l}^{i,j}}{\sum_{k=1}^{61} F_k^i N_k}.$$

We consider a hierarchical regression model that relates $x_{i,j}$ to $y_{i,j}$ and allows us to compare the expected and observed distances at the codon level for several properties simultaneously with the following rationale. If a given site i is neutral with respect to property j, then $y_{i,j} \approx x_{i,j}$. If property j is conserved at site i, then $y_{i,j} << x_{i,j}$ and finally, if property j is radically changing at site i, then $y_{i,j} >> x_{i,j}$. In addition, the model includes a latent factor structure for assessing the selective influence of several, possibly correlated, amino acid properties. We provide a detailed description of the model below.

2.2 The model

First we standardize $x_{i,j}$ and $y_{i,j}$ by dividing them by the maximum possible distance for each property and denote the standardized values as $x_{i,j}^*$ and $y_{i,j}^*$. This standardization will allow us to use a common prior on the model parameters associated with the amino acid properties. The first level of the model writes the expected value of the standardized observed distances as a simple linear function of the standardized expected distances, i.e., for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$ we have

$$y_{i,j}^* = \beta_{i,j} x_{i,j}^* + \epsilon_{i,j}, \tag{1}$$

with

$$\epsilon_{i,j} \sim \begin{cases} \mathsf{N}(0, \kappa^2) & \text{if } \beta_{i,j} = 0 \\ \mathsf{N}(0, \sigma^2/n_i^O) & \text{otherwise.} \end{cases}$$

The particular form for the error distribution is chosen to account for those $y_{i,j}^*$s that are equal to zero; indeed, some of the nonsynonymous changes can result in the same score values of the property being measured (for example, Asn, Asp, Glu and Gln all have the same score for hydropathy). We approximate a point mass at zero by assuming an appropriately tight prior for κ^2, which is the variance of the $y_{i,j}^*$s that are equal to zero. A different variance structure is imposed on the remaining $y_{i,j}^*$s. Further, since the number of observed nonsynonymous changes can be very different for different sites, $\mathrm{Var}(y_{i,j}^* | \beta_{i,j} \neq 0) = \sigma^2/n_i^O$, where n_i^O is the observed number of nonsynonymous changes at site i.

Our main focus is to model the latent structure underlying J amino acid properties while looking for evidence of selection at I sites using these properties. To that end, we use a factor structure (e.g., see Gorsuch [1983], Press [2005] and references therein) on the regression coefficients ($\beta_{i,j}$) in our model. In the simplest case, a K-factor ($K \leq J$) model on the i^{th} sample of a J-dimensional random quantity $\boldsymbol{\beta}_i$, is written as

$$\boldsymbol{\beta}_i = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\boldsymbol{f}_i + \boldsymbol{\nu}_i,$$

or, elementwise,

$$\beta_{i,j} = \alpha_j + \sum_{l=1}^{K} \lambda_{j,l} f_{i,l} + \nu_{i,j},$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)'$ is the mean vector; $\boldsymbol{\Lambda}$ is the $J \times K$ matrix of factor loadings; $\boldsymbol{f}_i = (f_{i,1}, \ldots, f_{i,K})'$ is the latent K-vector factor for the i^{th} sample and $\boldsymbol{\nu}_i = (\nu_{i,1}, \ldots, \nu_{i,J})'$ is the J-dimensional vector of independent, *idiosyncratic* ($\nu_{i,j}$ is unique to each $\beta_{i,j}$) noise terms, with $\boldsymbol{\nu}_i \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = \mathsf{diag}(\psi_1^2, \ldots, \psi_J^2)$. For identifiability, the factors $\boldsymbol{f}_i$ are traditionally assumed to be independently drawn from $\mathsf{N}(\mathbf{0}, \mathbf{I})$. Integrating out the latent factors in the model, we have $\text{Var}(\boldsymbol{\beta}_i|\boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ and $\text{Cov}(\beta_{i,j}, \beta_{i,j'}|\boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \sum_{l=1}^{K} \lambda_{j,l}\lambda_{j',l}$.

The K-factor model needs additional constraints to define an identifiable model. Firstly, $\boldsymbol{\Lambda}$, the matrix of factor loadings, has to be of full rank (K) to avoid identification problems arising from invariance of the model under location shifts [Geweke and Singleton, 1980]. Secondly, $\boldsymbol{\Lambda}$ needs a further constraint to avoid over-parameterization and thirdly, invariance under invertible linear transformations of the factors needs to be ensured. Following Lopes and West [2004], these constraints are satisfied by imposing the following structure on the loadings matrix,

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & 0 & 0 & \ldots & 0 \\ \lambda_{21} & \lambda_{22} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & 0 \\ \lambda_{K1} & \lambda_{K2} & \lambda_{K3} & \ldots & \lambda_{KK} \\ \lambda_{K+1,1} & \lambda_{K+1,2} & \lambda_{K+1,3} & \ldots & \lambda_{K+1,K} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \lambda_{J1} & \lambda_{J2} & \lambda_{J3} & \ldots & \lambda_{JK} \end{pmatrix}$$

where λ_{jj}, $j = 1, 2, \ldots, K$ are strictly positive. Due to the lower triangular structure of $\boldsymbol{\Lambda}$ the variability in the regression coefficient corresponding to the first property ($\beta_{i,1}$) is only determined by the first factor, while the second factor explains additional variability for all $\beta_{i,j}$s except $\beta_{i,1}$ and so on.

As pointed out earlier, we expect the $y^*_{i,j}$s to be zero or very close to zero for sites which strongly conserve a particular amino acid property. Since we are also interested in identifying such sites, we extend the structure on $\beta_{i,j}$ by adding a point mass at zero along with the factor structure when $\beta_{i,j}$ is different from 0, i.e.,

$$\beta_{i,j} \sim \pi_{i,0}\delta_0(\beta_{i,j}) + (1-\pi_{i,0})\mathsf{N}\Big(\beta_{i,j}|\alpha_j + \sum_{l=1}^{K}\lambda_{j,l}f_{i,l}, \psi_j^2\Big). \tag{2}$$

Here, $\delta_0(\cdot)$ is a point mass at zero and $\mathsf{N}(\cdot|m,v)$ is the Gaussian prior with mean m and variance v. The mixture probabilities ($\boldsymbol{\pi}$) are assumed to be site-specific but invariant across amino acid properties, whereas α_j and ψ_j^2 are assumed to be property-specific. In our previous analyses with protein-coding genes, we have seen that for most physicochemical properties a majority of the sites behave neutrally (i.e, $y_{i,j} \approx x_{i,j}$). For a sufficiently large sample size, i.e., for alignments with I large, we can consider a more general structure on $\beta_{i,j}$ with an additional point mass at $\beta_{i,j} = 1$ to model sites which are believed to be neutral with respect to a particular property [Datta et al., 2010]. We explore this scenario in more detail in Section 4.2.

Next we describe the priors which are essential to specify the full modeling structure in a Bayesian setting. For the factor loadings, priors of the form $\lambda_{j,l} \sim \mathsf{N}(0, C_0)$ when $j > l$; $l \leq K$ and $\lambda_{j,l} \sim \mathsf{N}(0, C_0)\mathbf{1}(\lambda_{j,l} > 0)$ for the diagonal elements $j = l = 1, \ldots, K$, will ensure that the K-factor model satisfies the identifiability constraints. The latter truncates the normal prior to restrict the diagonal elements to positive values. The upper diagonal elements are all 0, i.e., $\lambda_{j,l} = 0$ when $j < l$; $l \leq K$. However, in addition to explaining the correlations between the amino acid properties, we also want to be able to choose the number of factors required to explain the correlations. In order to achieve this, instead of using just normal densities and truncated normal densities as priors on the factor loadings we use sparsity-inducing priors [Lucas et al., 2006], i.e.,

$$\lambda_{j,l} \sim \gamma\delta_0(\lambda_{j,l}) + (1-\gamma)\mathsf{N}(\lambda_{j,l}|0, C_0)1_{\{\lambda_{j,l}>0\}}\ j = l = 1, \ldots, K, \tag{3}$$

$$\lambda_{j,l} \sim \gamma\delta_0(\lambda_{j,l}) + (1-\gamma)\mathsf{N}(\lambda_{j,l}|0, C_0)\ j > l, l \leq K. \tag{4}$$

The factors are assumed to have independent standard normal priors, i.e., $f_{i,l} \sim \mathsf{N}(0,1)$, while the ψ_j^2s are independently modeled as $\psi_j^2 \sim \mathsf{IG}(a_\psi, b_\psi)$. The remaining priors, i.e., priors on $\boldsymbol{\pi}$, α_j, γ and the scale parameters σ^2 and κ^2 are chosen to be conditionally conjugate to simplify calculations. In particular, the priors chosen are, $(\pi_{i,0}, 1-\pi_{i,0}) \sim \mathsf{Dir}(a_0, a_1)$, $\alpha_j \sim \mathsf{N}(1, C_\alpha)$, $\gamma \sim \mathsf{Beta}(a_\gamma, b_\gamma)$, $\kappa^2 \sim \mathsf{IG}(a_\kappa, b_\kappa)$ and $\sigma^2 \sim \mathsf{IG}(a_\sigma, b_\sigma)$. Note that in expectation all the properties are assumed neutral a priori, and thus

the α_js are centered around 1. Also note that in our definition of the inverse-gamma distribution IG(a, b), b corresponds to the scale (and not the rate) parameter of the distribution. Finally, $a_0, a_1, C_0, C_\alpha, a_\gamma, b_\gamma, a_\kappa, b_\kappa, a_\sigma, b_\sigma, a_\psi$, and b_ψ are assumed known. Section 4 illustrates how these hyperparameters are chosen in practice.

3 Posterior inference via MCMC methods

The structure of our model does not lead to closed forms for the posterior distributions of interest; hence, inference necessitates the use of Markov chain Monte Carlo (MCMC) methods. For a fixed K, Bayesian analyses of the latent factor model using MCMC methods are fairly straightforward. We introduce two sets of indicator variables to sample the mixture components for $\beta_{i,j}$ (equation 2) and $\lambda_{j,l}$ (equations 3 and 4). In particular, we introduce $\xi_{i,j} = 1_{\{\beta_{i,j} \neq 0\}}$ and $\theta_{j,l} = 1_{\{\lambda_{j,l} \neq 0\}}$. Due to our choice of priors, all the full conditionals have standard forms and sampling for all parameters proceeds via Gibbs steps. For more details about the MCMC algorithm and the actual forms of the full conditionals, please refer to Appendix 6.

The question remains of how to pick the number of factors K. In our applications, we start with a large number of factors and subsequently try to reduce such number. This is done by looking at the posterior means of the $\theta_{j,l}$s computed from the posterior samples $\theta_{j,l}^{(1)}, \ldots, \theta_{j,l}^{(B)}$, taken after MCMC convergence. The average $\bar{\theta}_{j,l} = \sum_{b=1}^{B} \theta_{j,l}^{(b)}/B$ gives an estimate of the posterior probability of the corresponding $\lambda_{j,l}$ being different from zero. We retain the l^{th} factor in our model if that factor has significant posterior probability of explaining variation in at least two variables, i.e., if $\bar{\theta}_{j,l} > c$ for some threshold c for at least two of the js. Following Lucas et al. [2006], we use $c = 0.95$. If one or more factors can be dropped based on the above criterion, we fit a model with a smaller number of factors and continue the above procedure until no more factors can be discarded.

4 Empirical exploration

We now present three simulation studies that aim to highlight some of the model features, illustrate the performance of the MCMC algorithms for posterior inference, and show how the inference on the latent factor structure is affected by the sample size and the proportion of sites that conserve or alter a given set of amino acid properties. For the first simulation study, we also provide a comparison with a model which assumes no correlations between variables.

Table 1: Posterior means of $\theta_{j,l}$s.

Variable	Factor 1	Factor 2	Factor 3	Factor 4
1	0.244	0.000	0.000	0.000
2	0.031	**1.000**	0.000	0.000
3	**1.000**	0.042	0.329	0.000
4	0.024	0.032	0.023	**1.000**
5	0.032	0.039	0.042	**1.000**
6	0.255	**1.000**	0.041	0.020
7	0.109	**1.000**	0.062	0.019
8	**1.000**	0.032	0.108	0.025
9	**1.000**	0.213	0.138	0.028

4.1 Simulation Study 1

Using a setup similar to Lopes and West [2004], a data set was simulated from a model with 100 observations (sites), 9 variables (properties), and 3 factors, i.e., $I = 100$, $J = 9$ and $K = 3$. In particular, $y_{i,j}$s were generated from equation (1) with $x_{i,j}$s $\sim \mathsf{Gamma}(3, 3)$. The $\beta_{i,j}$s needed in equation (1) were generated from equation (2) with $(\pi_{i,0}, 1 - \pi_{i,0}) \sim \mathsf{Dir}(2, 8)$, which implies that, in expectation, 20% of the sites strongly conserve the properties (i.e., 20% are expected to have $\beta_{i,j} = 0$) and 80% do not. The true factor loadings were set as follows:

$$\mathbf{\Lambda}' = \begin{pmatrix} 0.99 & 0 & 0 & 0.99 & 0.99 & 0 & 0 & 0 & 0 \\ 0 & 0.95 & 0 & 0 & 0 & 0.95 & 0.95 & 0 & 0 \\ 0 & 0 & 0.90 & 0 & 0 & 0 & 0 & 0.90 & 0.90 \end{pmatrix},$$

and $\mathbf{\Psi} = \mathsf{diag}(0.02, 0.19, 0.36, 0.02, 0.02, 0.19, 0.19, 0.36, 0.36)$. The true values for κ^2 and σ^2 were set at 0.001 and 0.1 respectively. The α_js were independently generated from a $\mathsf{N}(1, 0.25)$ distribution and the factors were generated from independent Gaussian distributions, i.e., $f_{i,l} \sim \mathsf{N}(0, 1)$ for all i and l.

MCMC analyses were performed under prior distributions with the following hyperparameters: $C_0 = 2$ (in Equations 3 and 4); $a_\psi = b_\psi = 2$ define fairly vague priors on ψ_j^2 with mean 0.5; $a_\kappa = 100$ and $b_\kappa = 10$ and $a_\sigma = 2$ and $b_\sigma = 10$, which define inverse gamma priors on κ^2 and σ^2 centered at approximately 0.001 and 0.1 respectively; α_j was given a Gaussian prior centered at 1 with variance $C_\alpha = 0.25$; $a_\gamma = 8$ and $b_\gamma = 2$, which define a beta prior with mean of 0.8; and finally, each pair $(\pi_{i,0}, 1 - \pi_{i,0})$ was given a Dirichlet prior with hyperparameters $a_0 = 1$ and $a_1 = 9$, which implies that in expectation, 10% of the sites are expected to be strongly conserved ($\beta_{i,j} = 0$) a priori. Posterior analyses are based on 1,000 samples taken after MCMC convergence. We discarded a burn-in of 10,000 iterations and

then used every 10^{th} iterate of another 10,000 iterations for a final sample of 1,000. We initially fitted a 4-factor model to the simulated data. Note that using our criterion for retaining factors (see Table 1), we can conclude that only 3 factors are required to describe the data. After fitting a 3-factor model with the priors specified above we obtained the following posterior summaries. The posterior mean and standard deviation of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$, up to two decimal places, were

$$\hat{\mathbf{\Lambda}}' = \begin{pmatrix} 1.01 & 0.01 & 0.01 & 0.99 & 0.97 & 0 & 0 & 0 & 0 \\ 0 & 0.98 & 0 & 0 & 0 & 1.04 & 0.98 & 0 & 0 \\ 0 & 0 & 1.06 & 0 & 0 & 0 & 0 & 0.90 & 0.93 \end{pmatrix}$$

with

$$\mathsf{sd}(\hat{\mathbf{\Lambda}}') = \begin{pmatrix} 0.10 & 0.05 & 0.06 & 0.08 & 0.08 & 0.02 & 0.02 & 0.01 & 0.02 \\ 0.00 & 0.08 & 0.03 & 0.02 & 0.01 & 0.13 & 0.11 & 0.03 & 0.03 \\ 0.00 & 0.00 & 0.12 & 0.01 & 0.02 & 0.11 & 0.03 & 0.15 & 0.13 \end{pmatrix}$$

and $\hat{\mathbf{\Psi}} = \mathsf{diag}(0.11,\ 0.18,\ 0.25,\ 0.09,\ 0.11,\ 0.33,\ 0.26,\ 0.40,\ 0.31)$ with $\mathsf{sd}(\mathsf{diag}(\hat{\mathbf{\Psi}})) = (0.04,\ 0.07,\ 0.11,\ 0.03,\ 0.04,\ 0.05,\ 0.05,\ 0.12,\ 0.09)$. Based on the posterior samples, we found that about 96% of the βs were classified correctly when the true β was different from 0, whereas 78% of the βs were classified correctly when the true β was 0, i.e., $\hat{Pr}(\hat{\beta}_{i,j} \neq 0 | \beta_{i,j} \neq 0, \mathbf{y}) = 0.96$ and $\hat{Pr}(\hat{\beta}_{i,j} = 0 | \beta_{i,j} = 0, \mathbf{y}) = 0.78$. In addition, $\hat{Pr}(\hat{\beta}_{i,j} = 0 | \beta_{i,j} \neq 0, \mathbf{y}) = 0.04$ and $\hat{Pr}(\hat{\beta}_{i,j} \neq 0 | \beta_{i,j} = 0, \mathbf{y}) = 0.22$. Figure 1 shows the true distribution and the histograms of the posterior samples for each of the α_js. The posterior samples of α_js all lie within 2 standard deviations of the true mean.

4.1.1 Comparison with results from a model with no factor structure

In this section, we investigate how the analysis is affected if we fit a model with no factor structure to the above simulated data. To that end, we modify (2) and fit a model where the regression coefficients $\beta_{i,j}$ are assumed to have the following structure:

$$\beta_{i,j} \sim \pi_{i,0}\delta_0(\beta_{i,j}) + (1 - \pi_{i,0})\mathsf{N}\Big(\beta_{i,j} | \alpha_j, \tau^2\Big). \tag{5}$$

This assumes that the $\beta_{i,j}$s come from a mixture with a point mass at zero, and a normal density otherwise. Note that in this case, $\mathrm{Cov}(\beta_{i,j}, \beta_{i,j'} | \boldsymbol{\alpha}, \tau^2) = 0$ and $\mathrm{Var}(\beta_{i,j} | \boldsymbol{\alpha}, \tau^2) = \tau^2$. Also, notice that in both (2) and (5), the prior mean on the $\beta_{i,j}$s different from 0 is α_j. Thus, we use the same priors for α_j and $(\pi_{i,0}, 1 - \pi_{i,0})$ as in the previous case, namely, $\mathsf{N}(1,\ C_\alpha = 0.25)$ and $\mathsf{Dir}(a_0 = 1, a_1 = 9)$. Since in (2), $\mathrm{Var}(\beta_{i,j} | \boldsymbol{\alpha}, \mathbf{\Lambda}, \mathbf{\Psi}) = \sum_{l=1}^{K} \lambda_{j,l}^2 + \psi_j^2$, the hyperparameters for the prior on τ^2 are chosen to agree with the priors induced

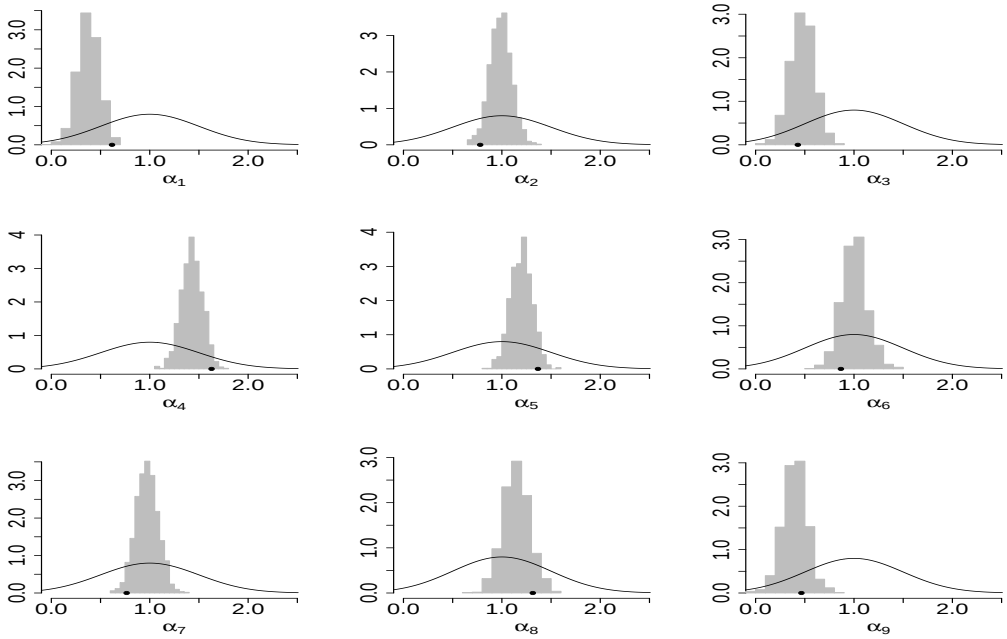

Figure 1: Histograms of the posterior samples of α_js with the $\mathsf{N}(1, 0.25)$ density, the distribution from which the α_js were generated. For this simulation, the prior distribution on α_j was also $\mathsf{N}(1, 0.25)$. The black dots represent the true values of α_j.

by the factor model in (2) as closely as possible, to make a fair comparison. In particular, we use an inverse gamma prior for τ^2 with hyperparameters $\mathsf{IG}(a_\tau = 2, b_\tau = 0.4)$, so that the prior mean on the variance parameter (τ^2) is 2.5. Posterior analyses are based on 1,000 samples after discarding the initial 10,000 as burn-in and using every 10^{th} iterate of the remaining 10,000 iterations.

The results from the analyses suggest that while both models have similar performances in terms of their ability to distinguish between βs from the different categories (0 or different) (see Table 2), the model with no factor structure, i.e., that using (5) as the prior for $\beta_{i,j}$, underestimates the correlation structure that is present in the data, as shown in Figure 2. The figure plots the histograms of the correlations between some of the pairs of variables for the regression model with factor structure (in light gray) and regression model (in dark gray). The regression model consistently underestimates the correlation between the pairs of variables, for example, among variables 1, 4 and 5. However, when there is actually no correlation between the variables, for example among variables 1, 2 and 3, both models estimate that correctly.

Table 2: Conditional probabilities of βs classified in each of the categories given the true categories for simulation 1, using a factor model and a regression model.

		Estimated			
		Factor		Regression	
		$\beta = 0$	β different	$\beta = 0$	β different
True	$\beta = 0$	**0.78**	0.22	**0.79**	0.21
	β different	0.04	**0.96**	0.03	**0.97**

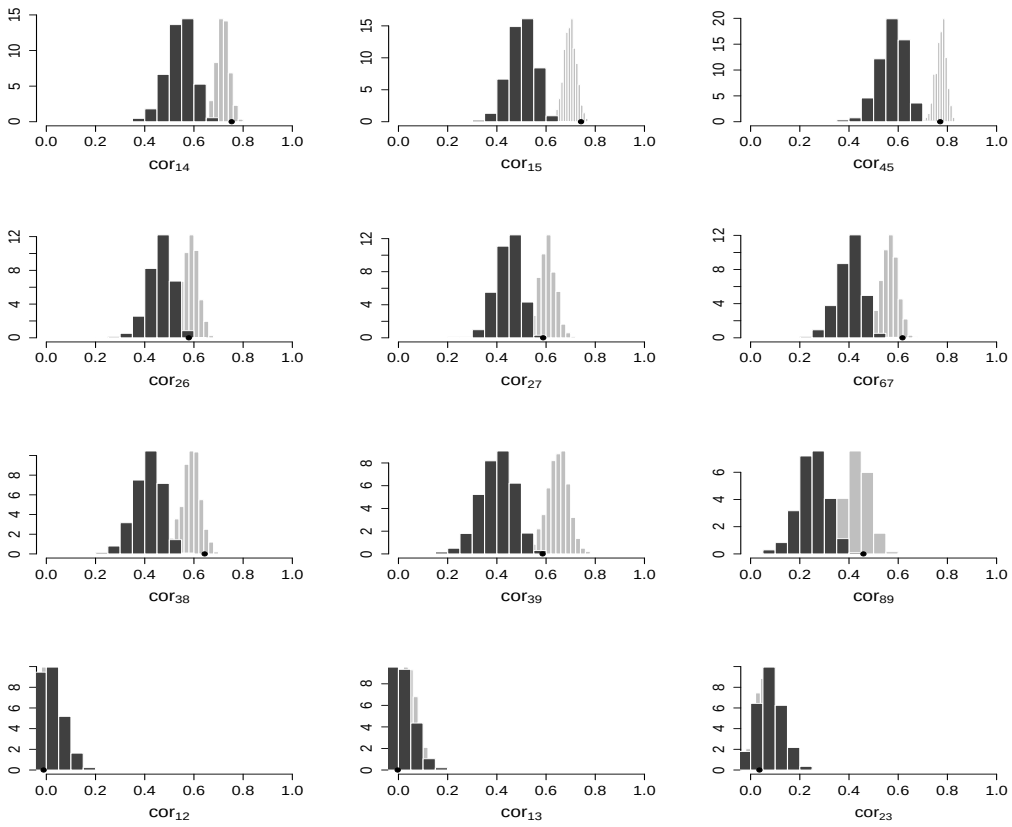

Figure 2: Histograms of the correlations between posterior samples of $\beta_{i,j}$ and $\beta_{i,j'}$ for different values of j and j' from the regression model with factor structure (light gray) and the regression model (dark gray). The black dots denote the true correlations between the respective $\beta_{i,j}$ and $\beta_{i,j'}$.

4.2 Simulation Study 2

As mentioned before, equation (2) can be modified to include an additional category with a point mass at $\beta_{i,j} = 1$ to account for sites that are neutral

with respect to a particular set of properties as done in Datta et al. [2010]. We present the results of a simulation study in which the data are sampled from a model in which the $\beta_{i,j}$s follow a mixture with three components. We highlight some model features and study the effect of the sample size on the estimation of the parameters of interest.

Once again, we assume that the model had 9 variables (properties). In this case we considered two different sample sizes, $I = 100$ and $I = 400$. For both sample sizes, our true model for each $\beta_{i,j}$ was a mixture with the following 3 components,

$$\beta_{i,j} \sim \pi_{i,0}\delta_0(\beta_{i,j}) + \pi_{i,1}\delta_1(\beta_{i,j}) + (1 - \pi_{i,0} - \pi_{i,1})\mathsf{N}\Big(\beta_{i,j}|\alpha_j + \sum_{l=1}^{3}\lambda_{j,l}f_{i,l}, \psi_j^2\Big), \quad (6)$$

with $(\pi_{i,0}, \pi_{i,1}, 1 - \pi_{i,0} - \pi_{i,1}) \sim \mathsf{Dir}(1, 1, 8)$, which implies that 80% of the sites are expected to have $\beta_{i,j} \neq 0, 1$, while 20% of the sites are expected to be strongly conserved ($\beta_{i,j} = 0$) or neutral with respect to property j. $\delta_1(.)$ denotes a point mass at 1. The true values for the factor loadings and the different variances were the same as those in the previous simulation and, as before, the factors were generated from independent $\mathsf{N}(0, 1)$ distributions. The α_js were independently generated from $\mathsf{N}(5, 1^2)$, $\mathsf{N}(1.8, 0.5^2)$ and $\mathsf{N}(0.4, 0.1^2)$, three from each distribution, to reflect three scenarios that we have encountered in practice, i.e., some sites may weakly conserve ($\mathsf{N}(0.4, 1^2)$), weakly alter ($\mathsf{N}(1.8, 0.5^2)$), or strongly alter ($\mathsf{N}(5, 1^2)$) a given property j.

The MCMC algorithm for the regression model with three components in the mixture for $\beta_{i,j}$ is detailed in Appendix 6. We initially fitted 4-factor models to the simulated data. The prior distributions used are the same as before, except the prior for the πs, since there are 3 components in π now. We used a $\mathsf{Dir}(1, 4, 1)$ as the prior for the πs, which implies that we expect about 66% of the sites to be neutral, 17% to be strongly conserved and 17% to be altered. As before, factors were retained in the model if a particular factor had significant (> 0.95) posterior probability of explaining the variation in at least two variables. For both sample sizes, using this criterion led to the conclusion that 3 factors were sufficient to explain the variability in the βs. The posterior summaries that result from fitting 3-factor models to the simulated data with $I = 100$ and $I = 400$ sites are discussed below.

100 observations

The posterior mean and standard deviation of the loadings $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ of the 3-factor model, to two decimal places, were

$$\hat{\mathbf{\Lambda}}' = \begin{pmatrix} 1.06 & 0.02 & 0.01 & 0.92 & 0.89 & 0 & 0 & 0 & 0 \\ 0 & 1.11 & 0.01 & 0 & 0 & 1.20 & 1.24 & 0.01 & 0 \\ 0 & 0 & 0.99 & 0 & 0 & 0.01 & 0 & 1.00 & 1.19 \end{pmatrix}$$

with

$$\mathsf{sd}(\hat{\mathbf{\Lambda}}') = \begin{pmatrix} 0.12 & 0.06 & 0.05 & 0.13 & 0.13 & 0.02 & 0.04 & 0.06 & 0.05 \\ 0.00 & 0.13 & 0.05 & 0.02 & 0.03 & 0.14 & 0.14 & 0.05 & 0.06 \\ 0.00 & 0.00 & 0.12 & 0.02 & 0.02 & 0.05 & 0.03 & 0.13 & 0.16 \end{pmatrix}$$

and $\hat{\mathbf{\Psi}} = \mathsf{diag}(0.12,\ 0.22,\ 0.17,\ 0.11,\ 0.10,\ 0.17,\ 0.15,\ 0.17,\ 0.26)$ with $\mathsf{sd}(\mathsf{diag}(\hat{\mathbf{\Psi}})) = (0.05,\ 0.08,\ 0.07,\ 0.04,\ 0.04,\ 0.07,\ 0.06,\ 0.08,\ 0.13)$.

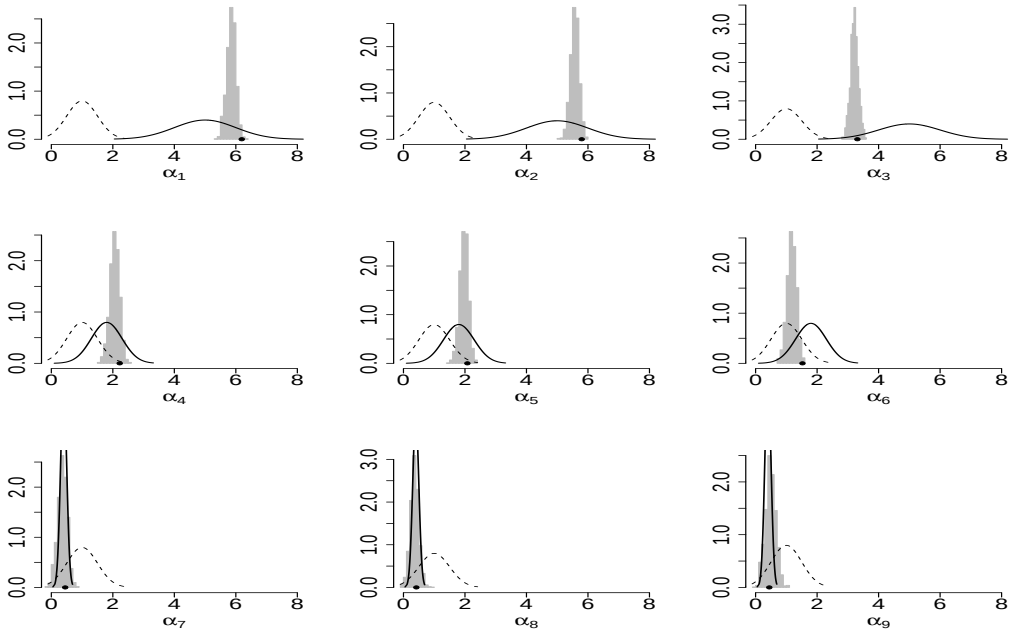

Figure 3: Histograms of the posterior samples of α_js with the true densities (in solid) and the prior density (dashed). The black dots denote the true α_j values.

Figure 3 shows the posterior samples of α_j for $j = 1, \ldots, 9$ along with the underlying true densities (in solid) and the prior (in dashed) used in the MCMC algorithm. In all cases, the model identifies the true structure of the α_js. Table 3 shows the conditional probabilities of βs which are classified as zero, one or different from zero and one given their true categories. About 29% of the $\beta_{i,j}$s that were sampled from the neutral category (i.e., those with $\beta_{i,j} = 1$) are misclassified as sites with $\beta_{i,j} \neq 0, 1$, while 90% of the sites sampled from the $\beta_{i,j} = 0$ and 78% of the sites from $\beta_{i,j} \neq 0, 1$ categories are correctly classified. Therefore, even though many of the posterior estimates based on data from only $I = 100$ sites are reasonable, it may be hard to distinguish between sites with $\beta_{i,j} = 1$ and sites whose $\beta_{i,j}$ values were simulated from the $\mathrm{N}(1.8, 0.5^2)$.

Table 3: Conditional probabilities of βs classified in each of the three categories given the true categories for simulation 2.

		Estimated					
		100 observations			400 observations		
		$\beta = 0$	$\beta = 1$	β different	$\beta = 0$	$\beta = 1$	β different
True	$\beta = 0$	**0.9**	0	0.1	**0.9**	0	0.1
	$\beta = 1$	0	**0.71**	0.29	0	**0.8**	0.2
	β different	0.02	0.20	**0.78**	0.02	0.19	**0.79**

400 observations

From the posterior analysis of the model with 3 factors the posterior mean and standard deviation of the loadings $\mathbf{\Lambda}$ and $\mathbf{\Psi}$, to two decimal places, were

$$\hat{\mathbf{\Lambda}}' = \begin{pmatrix} 0.92 & 0 & 0 & 0.93 & 0.94 & 0 & 0 & 0 & 0 \\ 0 & 0.98 & 0 & 0 & 0 & 1.05 & 0.96 & 0 & 0 \\ 0 & 0 & 0.75 & 0 & 0 & 0 & 0 & 0.97 & 1.03 \end{pmatrix}$$

with

$$\mathsf{sd}(\hat{\mathbf{\Lambda}}') = \begin{pmatrix} 0.04 & 0.01 & 0.01 & 0.05 & 0.05 & 0.01 & 0.01 & 0.02 & 0.01 \\ 0.00 & 0.05 & 0.01 & 0.00 & 0.01 & 0.06 & 0.05 & 0.02 & 0.08 \\ 0.00 & 0.00 & 0.06 & 0.01 & 0.02 & 0.03 & 0.02 & 0.08 & 0.08 \end{pmatrix}$$

and $\hat{\mathbf{\Psi}} = \mathsf{diag}(0.06,\ 0.21,\ 0.16,\ 0.05,\ 0.06,\ 0.14,\ 0.09,\ 0.44,\ 0.28)$ with $\mathsf{sd}(\mathsf{diag}(\hat{\mathbf{\Psi}})) = (0.01,\ 0.05,\ 0.05,\ 0.01,\ 0.01,\ 0.05,\ 0.03,\ 0.11,\ 0.10)$. We see that the $\hat{\mathbf{\Lambda}}'$ values are much closer to the true values for this example than for the previous case, indicating the number of observations is important, although for both cases the 95% credible intervals include the true values. From the posterior summaries of α_j, we find the histograms are more concentrated than in the previous simulation, which is expected since we have more observations in this case (figure not shown). Again, the posterior samples of all the α_js lie within 2 standard deviations of the true mean.

Table 3 shows the conditional probability of $\beta_{i,j}$s classified in the three categories given the true classes. As the sample size increases from $I = 100$ to $I = 400$, the model finds it easier to correctly categorize those sites with $\beta_{i,j} = 1$.

From the above simulation studies, the following points seem notable. In general, the model is flexible and can capture the underlying structure of the different parameters; for all cases, the 95% credible intervals included the true values of the parameters and, as expected, point estimates were closer to the true values with a larger sample size. In all the simulations presented

Table 4: Conditional probabilities of βs classified in each of the three categories, given the true categories. These results are based on a two-factor model, which is the model identified by our procedure.

		Estimated		
		$\beta = 0$	$\beta = 1$	β different
True	$\beta = 0$	**0.94**	0.06	0
	$\beta = 1$	0.01	**0.91**	0.08
	β different	0	0.29	**0.71**

here, more than 80% of the βs were correctly classified; however, identifying neutral sites when they are scarce and the sample size is not large enough can be difficult.

4.3 Simulation Study 3

In the simulations presented above about 80% of the βs were different from zero or one. Under this scenario, even the smallest of the sample sizes ($I = 100$) was sufficient to allow for accurate inference on the number of factors in the model and to correctly classify about 80% of the βs. However, when there is a smaller percentage of sites in the $\beta_{i,j} \neq 0, 1$ category, the model might be unable to identify the correct number of factors. We performed another simulation study where the data were sampled from a model with a smaller percentage of βs in the third category. Specifically, we set $I = 400$ and $J = 9$ as before and assumed that the $\beta_{i,j}$s were generated from equation (6) with $(\pi_{i,0}, \pi_{i,1}, 1 - \pi_{i,0} - \pi_{i,1}) \sim \mathsf{Dir}(1, 4, 1)$. This implies that approximately 16.7% of sites are expected to have $\beta_{i,j} \neq 0, 1$. All other components of the model were the same as in the previous simulation.

In this case, our approach to identifying the number of factors leads to selecting a model with only two factors. As shown in Table 4, this two-factor model finds it easier to categorize those βs for which the data provide a lot of information. In other words, large proportions of sites are correctly classified in the first and second categories (0.94 and 0.91, respectively) since approximately 66.7% of the sites had $\beta_{i,j} = 1$ and even though only about 16.7% of the sites had $\beta_{i,j} = 0$, such sites are relatively easy to distinguish from those with $\beta_{i,j} \neq 0, 1$. Contrast this with the results in Table 3, where there was a higher probability of correctly identifying sites with $\beta_{i,j} \neq 0, 1$ given their true category. In addition, we found that the ability to correctly classify the βs was not affected by the number of factors in the model, as similar results were seen if a 3-factor or a 4-factor model was fitted.

The insights provided by this simulation are important because in previous data analyses using models that do not incorporate the latent factor structure we have seen that there is usually a sizeable number of sites for

which $\beta_{i,j} = 1$ [Datta et al., 2010]. Therefore, when the focus is on discovering the latent factor structure underlying a large set of amino acid properties, as in our case, using a model with only two components in the mixture for $\beta_{i,j}$ — that pools sites with $\beta_{i,j} = 1$ with those for which $\beta_{i,j} \neq 0, 1$ — leads to increased information for estimating such structure. After identifying the factors, a model that includes only those amino acid properties that are roughly independent can then be used to determine which sites are altering or conserving such set of properties. If a site is conserving or altering a given set of properties it will likely have the same effect on other properties not included in the model but whose behavior can be explained by the set of independent properties via the latent factor structure. Thus, in the analysis presented below with lysin data, we consider a model with two categories for the $\beta_{i,j}$s (zero or different from zero and one) to first discover the latent structure across a relatively large set of amino acid properties. Then, once a much smaller set of roughly independent properties has been identified, one can fit a model with three categories on the $\beta_{i,j}$s in order to determine which sites are neutral, not altering, conserving or radically changing this set of properties as in [Datta et al., 2010].

5 Application to Lysin data

5.1 Data

Our proposed model was applied on the sperm lysin data set which consists of cDNA from 25 abalone species with 135 codons in each sequence [Yang et al., 2000b]. Sites with alignment gaps were removed from all sequences, which resulted in 122 codons for the analysis presented here. The phylogeny of Yang et al. [2000b] and the codon substitution model `M8` in `PAML`, version 3.15, [Yang, 1997] was used to generate the ancestral sequences. The model `M8` uses a discretized beta distribution to model ω values between zero and one with probability p_0 and allows for an additional positive selection category with $\omega > 1$ and probability p_1.

5.2 Analysis with amino acid properties

We analyzed the lysin data with 32 physicochemical properties (see Table 5). As mentioned in Section 2.1, we only look at nonsynonymous changes and for the lysin data set, this meant that our effective sample size was 94 codon sites. In order to be able to better estimate the factor structure, we assumed that the βs are modeled using the two-component mixture given in equation (2). The priors for the MCMC analyses were the same used in the first simulation (see Section 4.1). For the Gibbs sampler, we used a burn-in of 20,000 iterations and then saved every 10^{th} iterate of 100,000 iterations

Table 5: List of 32 amino acid properties used in the analysis.

Property	Symbol	Property	Symbol
Hydropathy	h	Helical contact area	C_a
Molecular volume	M_v	Isoelectric point	pH_i
Surrounding hydrophobicity	H_p	Long-range non-bonded energy	E_l
Polarity(Zimmerman)	p_{zim}	Mean r.m.s. fluctuation displacement	F
Alpha-helical tendencies	P_α	Molecular weight	M_w
Polarity(Grantham)	p	Normalized consensus hydrophobicity	H_{nc}
Average number of surrounding residues	N_s	Partial specific volume	V^0
Power to be at the C-terminal	α_c	Polar requirement	P_r
Composition	c	Power to be at the middle of alpha-helix	α_m
Compressibility	K^0	Power to be at the N-terminal	α_n
Equilibrium constant (ionization of COOH)	$pK^{'}$	Refractive index	μ
Beta-structure tendencies	P_β	Short and medium range non-bonded energy	E_{sm}
Bulkiness	B_l	Solvent accessible reduction ratio	R_a
Buriedness	B_r	Thermodynamic transfer hydrophobicity	H_t
Chromatographic index	R_F	Total non-bonded energy	E_t
Coil tendencies	P_c	Turn tendencies	P

for a final sample of 10,000. Examination of the trace plots from the MCMC output did not provide evidence of lack of MCMC convergence.

We started by fitting a model with 9 factors. Only 5 of the 9 factors had posterior probabilities greater than 0.95 of being different from 0 for at least two properties, which suggested that a model with 5 factors would suffice for this dataset. Following this, we fitted a 5 factor model to the data. An analysis of the posterior probabilities of the loadings being different from 0 revealed that we could not discard any more factors from the model. Figure 4 displays the percentages of variation in a property that is explained by each factor, which is calculated as $100 \times \frac{\lambda_{jl}^2}{\sum_{l=1}^{k} \lambda_{jl}^2 + \psi_j^2}$. We see that F1 explains most of the variation associated with properties h, H_p, B_r, F, R_a and E_t (about 69%-86%) and about 44% of the variation in H_{nc}. F2 explains most

of the variability in properties M_v, C_α, M_w, V^0, μ and E_{sm} (59%-76%). F3 explains 84%-87% variation in properties p_{zim} and pH_i and some variation in p, R_F, H_{nc} and P_r ($\approx$ 19%-53%). F4 explains additional variability in properties correlated with polarity, namely, p, N_s, P_β, B_l and R_F, while F5 explains variability in properties P_α, P_c and P (50%-60%). F5 also explains about 29% of the variability in α_m and α_n. A look into the correlation structure of the properties reveals that F1 explains variability in Hydropathy and the properties which are correlated with it. The same is true for the other factors (F2, ..., F5). The only two properties for which almost no variation is explained by this model are K^0 and pK'. These two properties are not correlated with any of the other properties or among each other. Since factors were retained only if the posterior probability of at least 2 properties being explained by a factor was greater than 0.95, the factors will not explain variation for independent variables. Five other properties for which less than 50% variation is explained by the factor structure are α_c, α_m, α_n, c and H_t (see Figure 5). For each of the remaining 25 properties, the factor structure explains more than 50% variability with the maximum being for p_{zim} (93%).

The model also allows us to draw conclusions about the average behavior of the different properties or about the behavior of specific sites. For example, Figure 6 shows the posterior samples of α_js of five of the properties which are representative of the five factors in the model and a sixth property for which the factor structure does not explain much of the variability. Some of the properties are mostly conserved like M_v or p, while h is mostly neutral and K^0 puts most of its mass above 1. Table 6 lists sites which maximize the posterior probability of the site being strongly conserved, i.e., $\Pr(\beta_{i,j} = 0|$ data) is maximized for these sites. A few of the properties for which no such sites were found are not reported in the table. Note that properties which are correlated with each other and can be explained by a common factor need not necessarily show the same behavior when it comes to specific sites. For example, p_{zim} and pH_i correlated and are explained by the same factor, but sites which maximize the probability that $\beta_{i,j}$ is strongly conserved are not the same for the 2 properties (sites 35, 101 and 117 for p_{zim} and site 101 for pH_i). Also in our model, the factor structure explains variability in $\beta_{i,j}$s when they are different from zero.

6 Discussion

We present a Bayesian hierarchical regression model with a latent factor structure that identifies radical, neutral or conserved amino acid changes by quantifying the magnitude of changes in amino acid properties. The latent factor structure in the model allows us to account for the correlation structure across a number of the physicochemical properties. The sparse

Table 6: Sites maximizing $\Pr(\beta_{ij} = 0|$ data) for lysin.

Property	Sites
	$\Pr(\beta_{ij} = 0|$ data)
H_p	27, 51, 68, 97, 117
p_{zim}	35, 101, 117
P_α	27, 68
N_s	51, 117
α_c	97
c	22, 28
K^0	35, 51, 117
pK^1	22, 28, 42, 97, 128
P_β	68, 97
B_l	35
P_c	22, 27, 57, 68
pH_i	101
E_l	19, 58, 105
F	97
H_{nc}	22
α_n	28, 68, 128
E_{sm}	19
R_a	27, 68
H_t	43, 51, 117
E_t	27, 68
P	97

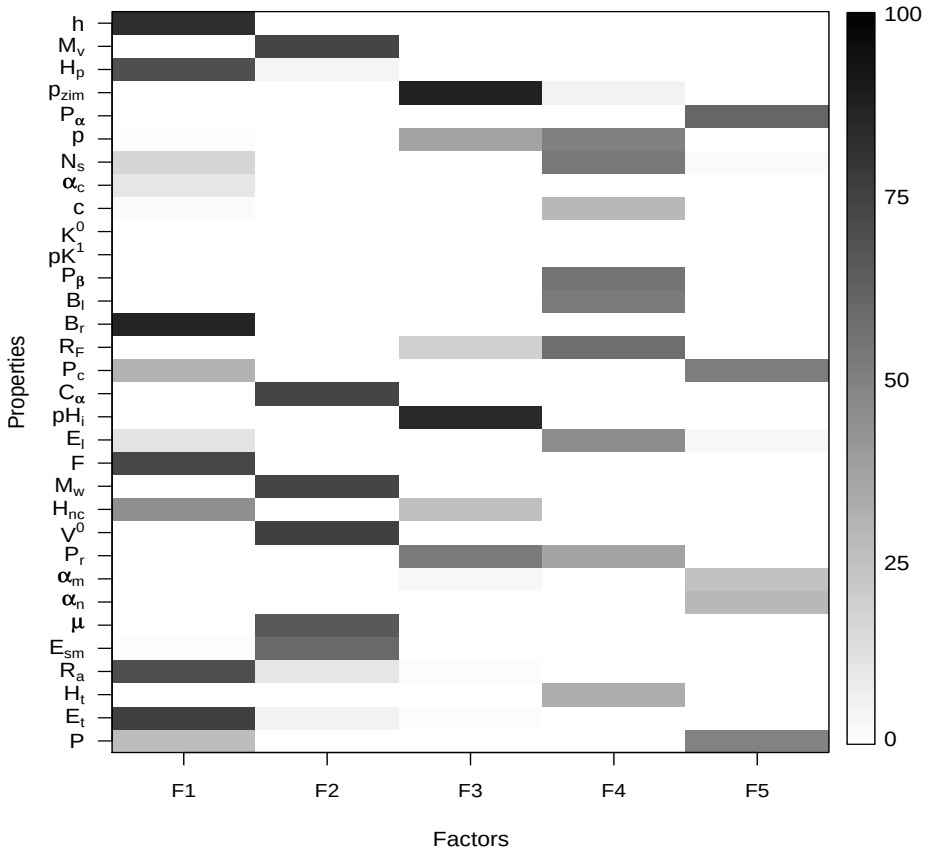

Figure 4: Percentages of variation in the 32 amino acid properties explained by 5 factors in lysin.

factor modeling structure also helps in reducing the dimension of the problem greatly and in simplifying the interpretation of the results. Another important feature of the model is its ability to provide site-specific as well as global results.

We considered two approaches for modeling the regression coefficients. One approach consists of using a 2-component mixture on the $\beta_{i,j}$s with a point mass at zero to account for sites that do not alter the j^{th} property, and a latent factor to describe the behavior of the coefficients for the remaining sites. The alternative approach uses a 3-component mixture on the $\beta_{i,j}$s that includes the two previous components plus a point mass at one to model neutral sites. We found that for a sufficiently large dataset and more importantly, with enough information to estimate the factor structure, it is possible to correctly identify the number of factors required in the model and to identify sites which strongly conserve or are neutral with respect to

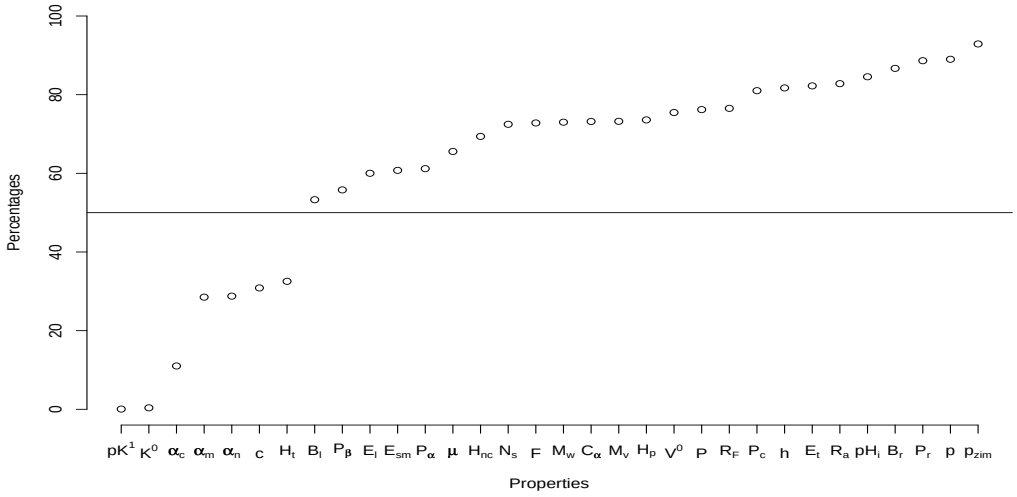

Figure 5: Total percentages of variation in the 32 amino acid properties explained by 5 factors in lysin. The horizontal line marks 50% on the y-axis.

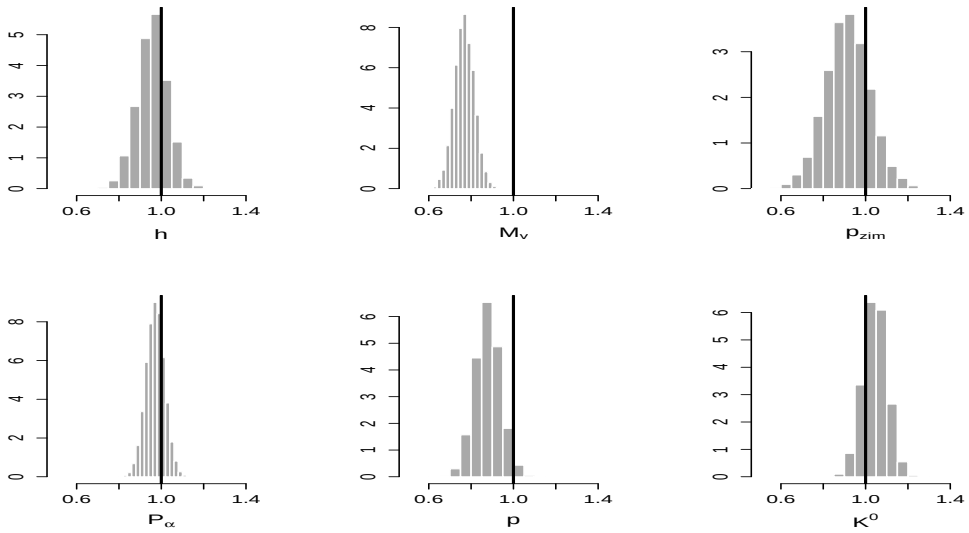

Figure 6: Histograms of the posterior samples of α_js of 6 properties for lysin. The first five properties are representative of the five factors in the factor model, while the sixth is one of the two properties for which the factor structure explains very little of the variability.

a given property. However, if there is reason to believe *a priori* that the percentage of sites to estimate the factor structure is relatively small, one should avoid using the model with the 3 categories ($\beta_{i,j} = 0, 1$ or different) directly.

In principle, the order of variables in a factor analysis could affect the posterior inference about the number of factors, but not the amount of variation being explained by the model [Lopes and West, 2004]. We avoid this issue by introducing sparsity-inducing priors on the factor loading. These sparsity-inducing priors give a positive probability to each of the factor loadings being equal to zero, enabling our procedure to select the number of factors to be fairly robust to the order of the properties, and allowing more flexibility in the identification of the factors. However, the order of the variables will still affect the interpretation of the factors and so, model users should be careful in selecting the order in which the properties will be included in the model. For instance, if one wants to interpret all the properties included in the model in terms of hydropathy and polarity — assuming that these are in fact significant properties — then they should be the first two properties listed so that they correspond to the first two factors. As a starting point, and in the absence of other prior information, users can choose some properties that are highly correlated with the remaining properties but not among each other as the first few properties. Users should also be reminded that the correlations among the properties calculated only from the scores of the 20 naturally occurring amino acids, obtained from the amino acid index database, may not be representative of the correlation structure present in the data, since the correlations in the data may be affected by other elements such as codon usage.

In the analyses presented here, we chose to begin fitting models with a relatively large number of factors and then decrease the number of factors as required. This idea is similar in spirit to backward elimination in stepwise regression. One can of course go the opposite route, i.e., start with a small number of factors and then increase the number as required, which is similar to the idea of forward selection. It might be argued that the initial fitting of the model with a large number of factors is computationally expensive, however, this approach will probably require less number of reruns to get to the correct number of factors. The choice of the maximum number of factors is a question that needs to be decided by the user, based on available computation power and other practical constraints.

Acknowledgements

Saheli Datta and Raquel Prado were supported by NIH/NIGMS grant R01GM072003-02, and Abel Rodríguez was supported by NIH/NIGMS grant R01GM090201-01.

References

Anisimova, M. and Kosiol, C. (2009). Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol. Biol. Evol.*, 26:255–271.

Atchley, W. R., Zhao, J., Fernandes, A. D., and Drüke, T. (2005). Solving the protein sequence metric problem. *P. Natl. Acad. Sci. USA*, 102:6395–6400.

Datta, S., Prado, R., Rodriguez, A., and Escalante, A. A. (2010). Characterizing molecular evolution: a hierarchical approach to assess selective influence of amino acid properties. *Bioinformatics*, 26:2818–2825.

Geweke, J. and Singleton, K. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *J. Amer. Stat. Assoc.*, 75:133 – 137.

Gorsuch, R. L. (1983). *Factor Analysis*. Psychology Press, 2nd edition.

Huelsenbeck, J. P. and Dyer, K. A. (2004). Bayesian estimation of positively selected sites. *J. Mol. Evol.*, 58:661–672.

Hughes, A., Ota, T., and Nei, M. (1990). Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Mol. Biol. Evol.*, 7:515–524.

Kidera, A., Konishi, Y., Oka, M., Ooi, T., and Scheraga, H. A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Protein Chem.*, 4:23–55.

Li, W. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.*, 36:96 – 99.

Lopes, H. and West, M. (2004). Bayesian model assessment in factor analysis. *Stat. Sinica*, 14:41 – 67.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006). Sparse statistical modelling in gene expression genomics. In Do, K. A., Müller, P., and Vannucci, M., editors, *Bayesian Inference for Gene Expression and Proteomics*, pages 155–176. Cambridge University Press, Cambridge, UK.

McClellan, D. and McCracken, K. (2001). Estimating the influence of selection on the variable amino acid sites of the cytochrome *b* protein functional domains. *Mol. Biol. Evol.*, 18:917 – 925.

McClellan, D., Palfreyman, E., Smith, M., Moss, J., Christensen, R., and Sailsbery, J. (2005). Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins. *Mol. Biol. Evol.*., 22:437–455.

Nei, M. and Kumar, S. (2000). *Molecular evolution and phylogenetics*. New York: Oxford University Press.

Press, S. J. (2005). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. Dover Publications, 2nd edition.

Ronquist, F. and Huelsenbeck, J. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572 – 1574.

Sainudiin, R., Wong, W. S. W., Yogeeswaran, K., Nasrallah, J. B., Yang, Z., and Nielsen, R. (2005). Detecting site-specific physicochemical selective pressures: Applications to the class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *J. Mol. Evol.*, 60:315–326.

Suzuki, Y. (2004). New methods for detecting positive selection at single amino acid sites. *J. Mol. Evol.*, 59:11–19.

Suzuki, Y. and Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.*, 16:1315–1328.

Wong, W. S. W., Sainudiin, R., and Nielsen, R. (2006). Identification of physicochemical selective pressure on protein encoding nucleotide sequences. *BMC Bioinformatics*, 7:148.

Xia, X. and Li, W. (1998). What amino acid properties affect protein evolution? *J. Mol. Evol.*, 47:557 – 564.

Yang, Z. (1997). Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.*, 42:294 – 307.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*, 19:908 – 917.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A. (2000a). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431 – 449.

Yang, Z. and Swanson, W. (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.*, 19:49 – 57.

Yang, Z., Swanson, W., and Vacquier, V. (2000b). Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineage and sites. *Mol. Biol. Evol.*, 17:1446–1455.

Zhang, J. (2000). Rates of conservative and radical non-synonymous nucleotide substitutions in mammalian nuclear genes. *J. Mol. Evol.*, 50:56–68.

Details about the MCMC

Consider the model presented in Section 2.2. As mentioned in Section 3, to sample the mixture components of $\beta_{i,j}$ and $\lambda_{j,l}$, we introduce two sets of indicators as below:

$$\xi_{ij} = \begin{cases} 0 & \text{if } \beta_{ij} = 0 \\ 1 & \text{otherwise.} \end{cases}$$

and

$$\theta_{jl} = \begin{cases} 0 & \text{if } \lambda_{jl} = 0 \\ 1 & \text{otherwise.} \end{cases}$$

The full conditionals for all the parameters are given below.

1. $p(\sigma^2|\cdots) \propto \prod_{\{(i,j)\,:\,\xi_{ij}\neq 0\}} \mathrm{N}(y^*_{ij}|\beta_{ij}x^*_{ij}, \sigma^2/n_i^O)\mathrm{IG}(\sigma^2|a_\sigma, b_\sigma)$ which yields

$$(\sigma^2|\cdots) \sim \mathrm{IG}\left[\frac{\sum_{ij} I_{\{\xi_{ij}\neq 0\}}}{2} + a_\sigma, \left(\frac{\sum_{\{(i,j)\,:\,\xi_{ij}\neq 0\}} n_i^O(y^*_{ij} - \beta_{ij}x^*_{ij})^2}{2} + \frac{1}{b_\sigma}\right)^{-1}\right]$$

2. $p(\kappa^2|\cdots) \propto \prod_{\{(i,j)\,:\,\xi_{ij}=0\}} \mathrm{N}(y^*_{ij}|\beta_{ij}x^*_{ij}, \kappa^2)\mathrm{IG}(\kappa^2|a_\kappa, b_\kappa)$ which yields $(\kappa^2|\cdots) \sim$

$$\mathrm{IG}\left[\frac{\sum_{ij} I_{\{\xi_{ij}=0\}}}{2} + a_\kappa, \left(\frac{\sum_{\{(i,j)\,:\,\xi_{ij}=0\}} (y^*_{ij} - \beta_{ij}x^*_{ij})^2}{2} + \frac{1}{b_\kappa}\right)^{-1}\right]$$

3. $p(\pi_{i0}, \pi_{i1}|\cdots) \propto \prod_{ij} \pi_{i0}^{I_{\{\xi_{ij}=0\}}} \pi_{i1}^{I_{\{\xi_{ij}=1\}}} \mathrm{Dir}(\pi_0, \pi_{i1}|a_0, a_1)$ resulting in $(\pi_{i0}, \pi_{i1}|\cdots) \sim$

$$\mathrm{Dir}\left(\sum_{j=1}^{J} I_{\{\xi_{ij}=0\}} + a_0, \sum_{j=1}^{J} I_{\{\xi_{ij}=1\}} + a_1\right)$$

4. $p(\alpha_j|\cdots) \propto \prod_{\{i\,:\,\xi_{ij}=1\}} \mathrm{N}(\beta_{ij}|\alpha_j + \sum_{l=1}^{k}\lambda_{jl}f_{il}, \psi_j^2)\mathrm{N}(\alpha_j|1, C_\alpha)$ leading to

   $(\alpha_j|\cdots) \sim \mathrm{N}\left[\frac{\alpha.\mathrm{mean}}{\alpha.\mathrm{var}}, \frac{1}{\alpha.\mathrm{var}}\right]$ where $\alpha.\mathrm{var} = \frac{1}{C_\alpha} + \sum_{\{i\,:\,\xi_{ij}=1\}} \frac{1}{\psi_j^2}$, $\alpha.\mathrm{mean} =$

   $$\frac{1}{C_\alpha} + \sum_{\{i\,:\,\xi_{ij}=1\}} \frac{(\beta_{ij} - \sum_{l=1}^{k}\lambda_{jl}f_{il})}{\psi_j^2}$$

5. $p(\psi_j^2|\cdots) \propto \prod_{\{i\,:\,\xi_{ij}=1\}} \mathrm{N}(\beta_{ij}|\alpha_j + \sum_{l=1}^{k}\lambda_{jl}f_{il}, \psi_j^2)\mathrm{IG}(\psi_j^2|a_\psi, b_\psi)$ which yields

   $$(\psi_j^2|\cdots) \sim \mathrm{IG}\left[\frac{\sum_i I_{\{\xi_{ij}=1\}}}{2} + a_\psi, \left(\frac{\sum_{\{i\,:\,\xi_{ij}=1\}} (\beta_{ij} - \alpha_j - \sum_{l=1}^{k}\lambda_{jl}f_{il})^2}{2} + \frac{1}{b_\psi}\right)^{-1}\right]$$

6. $p(f_{il}|\cdots) \propto \prod_{\{j\,:\,\xi_{ij}=1\}} \mathrm{N}(\beta_{ij}|\alpha_j + \sum_{l=1}^{k}\lambda_{jl}f_{il}, \psi_j^2)\mathrm{N}(f_{il}|0,1)$ yielding $(f_{il}|\cdots) \sim$

   $\mathrm{N}\left[\frac{f.\mathrm{mean}}{f.\mathrm{var}}, \frac{1}{f.\mathrm{var}}\right]$ where $f.\mathrm{var} = 1 + \sum_{\{j\,:\,\xi_{ij}=1\}} \frac{\lambda_{jl}^2}{\psi_j^2}$, $\quad f.\mathrm{mean} = \sum_{\{j\,:\,\xi_{ij}=1\}} \frac{\lambda_{jl}(\beta_{ij} - \alpha_j - \sum_{l' \neq l=1}^{k}$

7. θs are sampled by integrating over the corresponding λs.

   When $j \neq l$, $p(\theta_{jl}|\cdots) \sim \gamma \prod_{\{i\,:\,\xi_{ij}=1\}} \mathsf{N}(\beta_{ij}|\alpha_j + \sum_{l' \neq l=1}^{k}\lambda_{jl'}f_{il'}, \psi_j^2) +$

   $$(1-\gamma)\frac{1}{\sqrt{C_\lambda C_0^*}}\left(\prod_{\{i\,:\,\xi_{ij}=1\}} \frac{1}{\sqrt{2\pi\psi_j^2}}\right)\exp\left\{-\frac{1}{2}\left[\sum_{\{i\,:\,\xi_{ij}=1\}} \frac{z_{ij}^2}{\psi_j^2} - \frac{1}{C_0^*}\Big(\sum_{\{i\,:\,\xi_{ij}=1\}} \frac{z_{ij}f_{il}}{\psi_j^2}\Big)^2\right]\right\}$$

   where $C_0^* = \frac{1}{C_\lambda} + \frac{\sum_{\{i\,:\,\xi_{ij}=1\}} f_{il}^2}{\psi_j^2}$ and $z_{ij} = \beta_{ij} - (\alpha_j + \sum_{l' \neq l=1}^{k}\lambda_{jl'}f_{il'})$.

   For $j = l$, $p(\theta_{jl}|\cdots)$ has an expression similar to above, with the exception of additional normalizing constants, one from the prior and one from the corresponding updated truncated Gaussian distribution.

8. λs

   Diagonal λs (i.e., $j = l$)

- $(\lambda_{jl}|\cdots) = \begin{cases} 0 & \text{if } \theta_{jl}=0 \\ \mathrm{N}\left[\dfrac{\lambda.\mathrm{mean}}{\lambda.\mathrm{var}}, \dfrac{1}{\lambda.\mathrm{var}}\right] 1_{\{\lambda_{jl}>0\}} & \text{if } \theta_{jl}=1 \end{cases}$

  where $\lambda.\mathrm{var} = \dfrac{1}{C_\lambda} + \sum_{\{i\,:\,\xi_{ij}=1\}} \dfrac{f_{ij}^2}{\psi_j^2}$, $\lambda.\mathrm{mean} = \sum_{\{i\,:\,\xi_{ij}=1\}} \dfrac{f_{ij}(\beta_{ij} - \alpha_j - \sum_{l\neq j=1}^{k} \lambda_{jl} f_{il})}{\psi_j^2}$

  Non-diagonal λs (i.e., $j \neq l$)

  (a) $(\lambda_{jl}|\cdots) = \begin{cases} 0 & \text{if } \theta_{jl}=0 \\ \mathrm{N}\left[\dfrac{\lambda.\mathrm{mean}}{\lambda.\mathrm{var}}, \dfrac{1}{\lambda.\mathrm{var}}\right] & \text{if } \theta_{jl}=1 \end{cases}$

  where $\lambda.\mathrm{var} = \dfrac{1}{C_\lambda} + \sum_{\{i\,:\,\xi_{ij}=1\}} \dfrac{f_{il}^2}{\psi_j^2}$, $\lambda.\mathrm{mean} = \sum_{\{i\,:\,\xi_{ij}=1\}} \dfrac{f_{il}(\beta_{ij} - \alpha_j - \sum_{l'\neq l=1}^{k} \lambda_{jl'} f_{il'})}{\psi_j^2}$

9. $p(\gamma|\cdots) \propto \prod_{\{j\,:\,\xi_{ij}=1\}} \prod_{l=1}^{k} \gamma^{I_{\{\theta_{jl}=0\}}} (1-\gamma)^{I_{\{\theta_{jl}=1\}}} \mathrm{Beta}(\gamma|a_\gamma, b_\gamma)$ resulting in

   $(\gamma|\cdots) \sim \mathrm{Beta}\left(a_\gamma + \sum_{\{j\,:\,\xi_{ij}=1\}} \sum_{l=1}^{k} I_{\{\theta_{jl}=0\}}, b_\gamma + \sum_{\{j\,:\,\xi_{ij}=1\}} \sum_{l=1}^{k} I_{\{\theta_{jl}=1\}}\right)$

10. ξs are sampled by integrating over βs.

    $p(\xi_{ij}|\beta_{ij}-) \propto \mathrm{N}(y_{ij}^*|0,\kappa^2)\pi_{i0} + \mathrm{N}\left(y_{ij}^*|x_{ij}^*(\alpha_j + \sum_{l=1}^{k} \lambda_{jl} f_{il}), \dfrac{\sigma^2}{n_i^O} + x_{ij}^{*2}\psi_j^2\right)\pi_{i1}$

11. $(\beta_{ij}|\cdots) = \begin{cases} 0 & \text{if } \xi_{ij}=0 \\ \mathrm{N}\left[\dfrac{\beta.\mathrm{mean}}{\beta.\mathrm{var}}, \dfrac{1}{\beta.\mathrm{var}}\right] & \text{if } \xi_{ij}=1 \end{cases}$

    where $\beta.\mathrm{var} = \dfrac{1}{\psi_j^2} + \dfrac{n_i^O x_{ij}^{*2}}{\sigma^2}$, $\beta.\mathrm{mean} = \dfrac{\alpha_j + \sum_{l=1}^{k} \lambda_{jl} f_{il}}{\psi_j^2} + \dfrac{n_i^O y_{ij}^* x_{ij}^*}{\sigma^2}$.

    NOTE:
    If instead of equation (2), the structure on $\beta_{i,j}$ is

$$\beta_{ij} \sim \pi_{i,0}\delta_0(\beta_{ij}) + \pi_{i,1}\delta_1(\beta_{ij}) + \pi_{i,2} N(\beta_{ij}|\alpha_j + \sum_{l=1}^{k} \lambda_{jl} f_{il}, \psi_j^2), \pi_{i,2} = 1 - \pi_{i,0} - \pi_{i,1},$$

we use a Dirichlet distribution with 3 components as the prior for π i.e., $(\pi_{i0}, \pi_{i1}, \pi_{i2}) \sim \mathsf{Dir}(a_0, a_1, a_2)$ and the following changes are necessary to the algorithm for generating posterior samples.

1. The dummy variables ξ_{ij} have to be defined such that

$$\xi_{ij} = \begin{cases} 0 & \text{if } \beta_{ij} = 0 \\ 1 & \text{if } \beta_{ij} = 1 \\ 2 & \text{otherwise.} \end{cases}$$

2. The posterior distribution for πs is given by

$$(\pi_{i0}, \pi_{i1}, \pi_{i2}| \cdots) \sim \text{Dir}\left(\sum_{j=1}^{J} I_{\{\xi_{ij}=0\}} + a_0, \sum_{j=1}^{J} I_{\{\xi_{ij}=1\}} + a_1, \sum_{j=1}^{J} I_{\{\xi_{ij}=2\}} + a_2\right)$$

3. $\xi_{i,j}$s are sampled as below

$$p(\xi_{ij}|\beta_{ij}-) \propto \mathsf{N}(y^*_{ij}|0, \kappa^2)\pi_{i0} + \mathsf{N}(y^*_{ij}|x^*_{ij}, \frac{\sigma^2}{n_i^O})\pi_{i1} + \mathsf{N}\left(y^*_{ij}|x^*_{ij}(\alpha_j + \sum_{l=1}^{k} \lambda_{jl} f_{il}), \frac{\sigma^2}{n_i^O} + x^{*2}_{ij}\psi_j^2\right)\pi_{i2}$$

4. $$(\beta_{ij}| \cdots) = \begin{cases} 0 & \text{if } \xi_{ij} = 0 \\ 1 & \text{if } \xi_{ij} = 1 \\ \text{N}\left[\dfrac{\beta.\text{mean}}{\beta.\text{var}}, \dfrac{1}{\beta.\text{var}}\right] & \text{if } \xi_{ij} = 2 \end{cases}$$

where $\beta.\text{var} = \dfrac{1}{\psi_j^2} + \dfrac{n_i^O x^{*2}_{ij}}{\sigma^2}$, $\beta.\text{mean} = \dfrac{\alpha_j + \sum_{l=1}^{k} \lambda_{jl} f_{il}}{\psi_j^2} + \dfrac{n_i^O y^*_{ij} x^*_{ij}}{\sigma^2}$.