The Bayesian Group-Lasso for Analyzing Contingency Tables

Sudhir Raman SUDHIR.RAMAN@UNIBAS.CH
Department of Computer Science, University of Basel, Bernoullistr. 16, CH-4056 Basel, Switzerland

Thomas J. Fuchs THOMAS.FUCHS@INF.ETHZ.CH
Institute of Computational Science, ETH Zurich, Universitaetstrasse 6, CH-8092 Zurich, Switzerland & Competence Center for Systems Physiology and Metabolic Diseases - Schafmattstr. 18, CH-8093 Zurich, Switzerland

Peter J. Wild PETER.WILD@CELL.BIOL.ETHZ.CH
Institute of Pathology, University Hospital Zurich, Schmelzbergstrasse 12, CH-8091 Zurich, Switzerland

Edgar Dahl EDAHL@UKAACHEN.DE
Institute of Pathology, University Hospital, Pauwelsstrasse 30, 52074 Aachen, Germany

Volker Roth VOLKER.ROTH@UNIBAS.CH
Department of Computer Science, University of Basel, Bernoullistr. 16, CH-4056 Basel, Switzerland

Abstract

Group-Lasso estimators, useful in many applications, suffer from lack of meaningful variance estimates for regression coefficients. To overcome such problems, we propose a full Bayesian treatment of the Group-Lasso, extending the standard Bayesian Lasso, using hierarchical expansion. The method is then applied to Poisson models for contingency tables using a highly efficient MCMC algorithm. The simulated experiments validate the performance of this method on artificial datasets with known ground-truth. When applied to a breast cancer dataset, the method demonstrates the capability of identifying the differences in interactions patterns of marker proteins between different patient groups.

1. Introduction and Related Work

The identification of important explanatory factors for a process is a key task in many practical learning problems. In the context of standard linear regression, the Lasso (Tibshirani, 1996) has become a popular method for this purpose in the recent years. The Lasso optimizes a regression functional under an ℓ_1-constraint on the coefficient vector: given a $n \times 1$ vector of responses $\boldsymbol{y} = (y_1, \ldots, y_n)^t,\ y_i \in \mathbb{R}$ and observation vectors $\boldsymbol{x}_i \in \mathbb{R}^p$ arranged as rows of a $n \times p$ data matrix X, the Lasso minimizes

$$\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_1 \leq \kappa. \tag{1}$$

In many applications, however, explanatory factors do not necessarily have a one-to-one correspondence with single features (main effects) but rather require a more complex representation, for instance, by considering not only single features but higher order interactions between some/all features. As a further extension to this idea, the factors (main effects + higher order interactions) may need to be represented as *groups* of variables. A popular example of this kind is the representation of categorical variables (i.e. "factors" in the usual statistical terminology and/or their interactions) as groups of "dummy" variables.

To address such situations, the Group-Lasso (Yuan & Lin, 2006) is an ideal choice since it serves as a natural extension of the Lasso by finding solutions that are sparse at the level of *groups* of variables which, for instance, might represent categorical variables and/or interaction terms. Despite the fact that the Group-Lasso estimators have proven useful in many applications (Kim et al., 2006; Meier et al., 2008; Roth & Fischer, 2008), their main problem concerns the definition of meaningful variance estimates for the regression coefficients. This problem because the Hessian is not defined at the optimal solution. In this paper, we suggest a full Bayesian treatment of the Group-Lasso to overcome this problem. The proposed model uses a hierarchical expansion and directly extends corresponding Bayesian versions of the standard Lasso (Figueiredo & Jain, 2001; Park & Casella, 2008) to handle grouped predictors.

Appearing in *Proceedings of the 26^{th} International Conference on Machine Learning*, Montreal, Canada, 2009.

While the Bayesian Group-Lasso is applicable for many different likelihood models, the focus of this paper is on Poisson models for contingency tables. In applications of this kind, the "dummy" covariates $\boldsymbol{x}$ represent factor interactions up to a certain order, and sparsity induced by the Group-Lasso corresponds to selecting edges in the hypergraph defining the interaction structure among these factors. Implementation approaches include the method described in (Figueiredo & Jain, 2001) which basically reduces to MAP point estimation without essential Bayesian features such as posterior variances, whereas (Park & Casella, 2008) describes a full Bayesian approach based on MCMC sampling. A third class of methods falls in between these extremes by utilizing approximation schemes like expectation propagation (Seeger, 2008). By exploiting a special orthogonality property of design matrices in the Poisson model for contingency tables, we are able to derive a highly efficient MCMC sampling algorithm that makes it possible to follow the full Bayesian route without resorting to approximation schemes even in large real-world examples. The availability of posterior variances and correlations overcomes several problems of a related penalized likelihood (or MAP) approach introduced in (Dahinden et al., 2007). Moreover, our algorithm is applicable to contingency tables stemming from arbitrary factors rather than being restricted to binary variables. We show that our method is capable of identifying relevant high-order interaction patterns both in toy examples and in large-scale medical applications.

2. The Bayesian Group-Lasso

For the following derivation, it is convenient to partition the data matrix X, and the coefficients $\boldsymbol{\beta}$ into G subgroups: $X = (X_1, \ldots, X_G)$,

$$\boldsymbol{\beta}^t = \left(\boldsymbol{\beta}_1^t, \ldots, \boldsymbol{\beta}_G^t\right). \tag{2}$$

The size of the g-th subgroup will be denoted by p_g.

The Group-Lasso minimizes the negative log-likelihood viewed as a function in $\boldsymbol{\beta}$, $l = l(\boldsymbol{\beta})$, under a constraint on the sum of the ℓ_2-norms of the subvectors $\boldsymbol{\beta}_g$:

$$\text{minimize } l(\boldsymbol{\beta}) \quad \text{s.t.} \quad \textstyle\sum_{g=1}^{G} \|\boldsymbol{\beta}_g\| \leq \kappa. \tag{3}$$

From a probabilistic perspective, the Group-Lasso with Gaussian likelihood can be understood as a standard linear regression model with Gaussian noise and a product of multivariate Laplacian priors over the regression coefficients. The observations $\boldsymbol{y} = (y_1, \ldots, y_n)^t$ are assumed to be generated according to

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + \varepsilon_i, \quad \text{with} \quad \varepsilon_i \sim N(0, \sigma^2), \tag{4}$$

which implies a likelihood of the form

$$\begin{aligned} p(\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2) &\propto \exp\left\{-\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2/(2\sigma^2)\right\} \\ &\propto (\sigma^2)^{-p/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^t X^t X (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right\} \\ &\quad \cdot (\sigma^2)^{-\nu/2} \exp\left\{-\frac{SSE}{2\sigma^2}\right\}, \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{\beta}}$ is the least-squares solution, $SSE = (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})$ is the sum of squared errors, and $\nu = n - p$. The last equation results from "completing the squares" which is standard in Bayesian regression, see for instance (Gelman et al., 1995). Assuming a multivariate (spherical) p_g-dimensional Multi-Laplacian prior over each group of regression coefficients,

$$\text{M-Laplace}(\boldsymbol{\beta}_g|\mathbf{0}, c^{-1}) \propto c^{p_g/2} \exp(-c\|\boldsymbol{\beta}_g\|_2), \tag{6}$$

the classical group-Lasso in eq. (3) is recovered as the MAP-solution in log-space, with $\sigma^2 c$ having the role of a *fixed* Lagrange parameter. For a full Bayesian treatment, on the other hand, we would like to place (hyper)priors over c and σ^2 and integrate out these parameters. In practice, these integrations will not be possible analytically, and we propose to use a Gibbs sampling strategy for stochastic integration.

Hierarchical expansion. For finding a representation in which all conditionals have a standard form, we use the following hierarchical expansion which extends the respective derivation for the standard Lasso in (Figueiredo & Jain, 2001) to grouped predictors: the prior can be expressed as a two-level hierarchical model involving independent zero-mean Gaussian priors over $\boldsymbol{\beta}_g$ and Gamma priors over λ_g. Defining $a_g = p_g\rho$ and $b_g = \|\boldsymbol{\beta}_g\|_2^2/\sigma^2$ (for each group g), the Multi-Laplacian prior on $\boldsymbol{\beta}_g$ can be expressed as a hierarchical Normal-Gamma model:

$$\begin{aligned} p(\boldsymbol{\beta}_g|\rho) &= \textstyle\int_0^\infty N(\boldsymbol{\beta}_g|0, \sigma^2\lambda_g^2)\text{Gamma}(\lambda_g^2|\frac{p_g+1}{2}, \frac{2}{a_g})\, d\lambda_g^2 \\ &= (\sigma^2)^{-\frac{p_g}{2}} \int_0^\infty \underbrace{(\lambda_g^2)^{-\frac{1}{2}} \exp\left[-\frac{b_g}{2\lambda_g^2} - \lambda_g^2 \frac{a_g}{2}\right]}_{(\frac{b_g}{a_g})^{\frac{1}{4}} K_{\frac{1}{2}}\left[(a_g b_g)^{\frac{1}{2}}\right] \cdot \text{GIG}(\lambda_g^2|\frac{1}{2}, a_g, b_g)} a_g^{\frac{p_g+1}{2}} d\lambda_g^2 \\ &\propto (\sigma^2)^{-\frac{p_g}{2}} (\tfrac{b_g}{a_g})^{\frac{1}{4}} (a_g b_g)^{-\frac{1}{4}} a_g^{\frac{p_g+1}{2}} \exp\left[-(a_g b_g)^{\frac{1}{2}}\right] \\ &= (a_g/\sigma^2)^{\frac{p_g}{2}} \exp(-(a_g/\sigma^2)^{\frac{1}{2}}\|\beta_g\|_2) \\ &\propto \text{M-Laplace}(\boldsymbol{\beta}_g|\mathbf{0}, (a_g/\sigma^2)^{-\frac{1}{2}}). \end{aligned} \tag{7}$$

In the above derivation we have used the definition of the generalized inverse Gaussian distribution

$$\begin{aligned} \text{GIG}\left(\lambda_g^2|\tfrac{1}{2}, a_g, b_g\right) &= (\tfrac{b_g}{a_g})^{-\frac{1}{4}} K_{\frac{1}{2}}^{-1}\left[(a_g b_g)^{\frac{1}{2}}\right] \\ &\quad \cdot (\lambda_g^2)^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}\left(\tfrac{b_g}{\lambda_g^2} + \lambda_g^2 a_g\right)\right], \end{aligned} \tag{8}$$

with $K_{\frac{1}{2}}(x) = \sqrt{\pi/(2x)}\exp[-x]$ denoting a special case of the spherical Bessel functions.

For the standard linear model we can, thus, "expand" the Group-Lasso in terms of a Gaussian likelihood, Gaussian priors over the regression coefficients and Gamma priors over the corresponding variances. The model is completed by introducing a prior on σ^2 (we use the standard conjugate joint prior $p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) = N(\boldsymbol{\beta}|\boldsymbol{\mu}, \sigma^2\Sigma) \cdot$ Inv-$\chi^2(\sigma^2|\nu_0, s_0^2)$) and a conjugate Gamma$(\rho|r, s)$ prior on ρ.

Multiplying the priors with the likelihood and rearranging the relevant terms yields the full conditional posteriors, which are needed in the Gibbs sampler for carrying out the stochastic integrations. Concerning $\boldsymbol{\beta}$ and σ^2, the resulting conditionals have the standard form in Bayesian regression: $p(\sigma^2|\bullet) =$ Inv-χ^2, and

$$\begin{aligned} p(\boldsymbol{\beta}|\bullet) =& N(\boldsymbol{\beta}|\tilde{\boldsymbol{\mu}}, \sigma^2\widetilde{\Sigma}) \text{ with } \widetilde{\Sigma} = (X^tX + \Lambda^{-1})^{-1}, \\ & \text{and } \tilde{\boldsymbol{\mu}} = \widetilde{\Sigma}X^tX\hat{\boldsymbol{\beta}}, \end{aligned} \tag{9}$$

where Λ is a diagonal matrix consisting of λ_g^2's as diagonal elements, with each λ_g^2 repeated p_g times. The conditional posterior of ρ is again Gamma due to conjugacy,

$$p(\rho|\bullet) = \rho^{Gr}\exp\left[-\rho\left(\tfrac{G}{s} + \tfrac{1}{2}\textstyle\sum_{g=1}^{G}\lambda_g^2 p_g\right)\right], \tag{10}$$

where r and s are the shape- and scale hyperparameters of the Gamma prior on ρ. Finally, the conditional of λ_g^2 is generalized inverse Gaussian:

$$p(\lambda_g^2|\bullet) = \text{GIG}\left(\lambda_g^2|\tfrac{1}{2}, a_g, b_g\right) \tag{11}$$

3. Poisson Models for Contingency Tables

While in principle the Bayesian Group-Lasso introduced in the last section is applicable for many different likelihoods, in this paper we focus on Poisson models for analyzing count data in contingency tables. The reason for this choice is threefold: (i) count data for certain compositions of discrete properties occur frequently in practical applications, particularly in a bio-medical context; (ii) feature selection for categorical variables is directly related to methods inferring sparsity on the level of groups of predictors; (iii) the use of certain encoding schemes for categorical variables allows us to derive a highly efficient sampling algorithm that makes full Bayesian inference practical in large-scale applications.

Denote by $\boldsymbol{z} = (z_1, \ldots, z_n)$ the observed counts in a contingency table with n cells. The standard approach to modeling count data is Poisson regression which involves a log-linear model with independent terms: for $i = 1, \ldots, n$:

$$z_i|\mu_i \sim \text{Poisson}(\mu_i) = \frac{\mu_i^{z_i}e^{-\mu_i}}{z_i!}, \tag{12}$$

with the Poisson mean $\mu_i = e^{\eta_i}$, and $\eta_i = \boldsymbol{x}_i\boldsymbol{\beta}$. Using a *random link*

$$\eta_i = \boldsymbol{x}_i^t\boldsymbol{\beta} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2) \tag{13}$$

makes it possible to allow for deviations from the log-linear model. For instance, overdispersed Poisson models (which are common in practice) can be modeled with random link functions. In the Bayesian context such random links correspond to a conditional Gaussian prior on η:

$$p(\eta_i|\boldsymbol{\beta}, \sigma^2) = N(\eta_i|\boldsymbol{x}_i^t\boldsymbol{\beta}, \sigma^2). \tag{14}$$

Note that inputs are available only as counts $\boldsymbol{z}$. Assuming that the total number of counts is fixed, the observed vector of individual counts $\boldsymbol{z} = (z_1, \ldots, z_n)$ can be considered as a realization drawn from a multinomial distribution. This is the approach taken in (Dahinden et al., 2007) in a penalized likelihood framework. If, on the other hand, we assume a sampling model in which the total number of counts itself is random and the time period of observing the counts is fixed, we arrive at the Poisson model (12). This sampling model is plausible for many practical situations. For example, a clinical study of fixed length where the counts correspond to the number of patients with certain properties visiting the hospital. The main technical advantage of the Poisson model lies in the factorization over the cells, given the means μ_i (see eq. (12)). To understand the nature of the "dummy" covariates $\boldsymbol{x}$, the following notations are helpful: assume our contingency table is based on observations of d categorical random variables or factors, $C_1, ..., C_d$, where each C_j has K_j levels. The "dummy" covariates $\boldsymbol{x}$ represent factor interactions, starting with the interactions of order zero (the main effects) up to all interactions of a certain maximum order Q. Interactions are denoted by the colon operator (:). There is one group of dummy variables used for encoding each interaction term. This corresponds to assuming that the means of the η_i (which are in turn the log Poisson means) can be expressed (in vector form) as $\bar{\boldsymbol{\eta}} = X\boldsymbol{\beta}$, with the design matrix X composed of individual submatrices:

$$\begin{aligned} X = [\mathbf{1}, & \underbrace{X^{C_1}, \ldots, X^{C_d}}_{\text{main effects}}, \underbrace{X^{C_1:C_2}, \ldots, X^{C_{d-1}:C_d}}_{\text{1st order interactions}}, \\ & \ldots, \underbrace{X^{C_1:\cdots:C_{Q+1}}, \ldots, X^{C_{d-Q}:\cdots:C_d}}_{\text{highest order interactions}}]. \end{aligned} \tag{15}$$

To avoid over-parametrization, identifiability constraints are imposed on the individual submatrices in the form of *contrast codes* which encode a factor with K levels into a matrix with $K - 1$ columns. Interaction terms are basically column-wise product expansions of these individual (main-effect) matrices. In many practical applications we are given *ordered* factors, i.e. ordinal variables for which a natural ordering is involved. Examples of this kind are

for instance intensity levels. For such ordinal categorical variables, the use of *polynomial contrast codes* is a natural choice. These encodings employ orthogonal polynomials and have the practical advantage that the (typically huge) design matrix is orthogonal, i.e. $X^t X = I$.

The use of random links not only allows for deviations from the parametric model, but also greatly simplifies Gibbs sampling: when conditioning on η_i, sampling of $\boldsymbol{\beta}$ and σ follows the standard procedure in Bayesian regression. The orthogonality property of the design matrix, $X^t X = I$, makes it possible to sample from very high-dimensional models in an efficient and numerically stable way since only diagonal matrices have to be inverted, see eq. (9). Updating η_i, on the other hand, is more difficult because the corresponding conditional is not of recognized form:

$$p(\eta_i|\boldsymbol{\beta}, \sigma^2, X, \boldsymbol{z}) \propto \exp\big[\sum_i \eta_i z_i - \exp(\eta_i) - \frac{1}{2\sigma^2}(\boldsymbol{x}_i^t\boldsymbol{\beta} - \eta_i)^2\big]. \tag{16}$$

However, the above conditional posterior is log-concave which makes it possible to use "black-box" sampling methods like adaptive rejection sampling. Alternatively, we propose to use a Laplace approximation similar to that in (Green & Park, 2004), which in practice turns out to give results which are almost indistinguishable from adaptive rejection sampling while speeding up the computations considerably. Figure 1 summarizes the hierarchical structure of the Poisson Group-Lasso model.

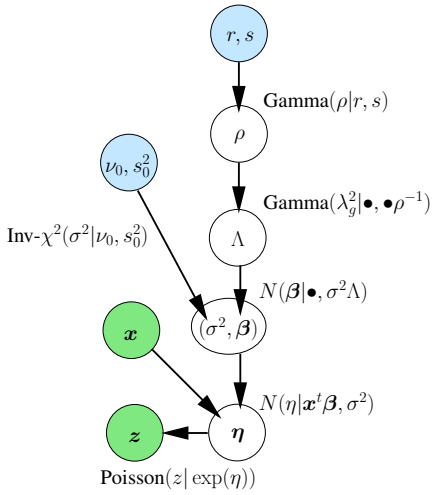

Figure 1. Dependency structure of the hierarchical Poisson Group-Lasso model. On top of this hierarchy are the hyperparameters (r, s) for the Gamma prior on ρ and (ν_0, s_0^2) for the Inv-χ^2 prior on σ^2. On the other end there are the observed vector of counts $\boldsymbol{z}$ and the "dummy" covariates $\boldsymbol{x}$. The core of the diagram represents the hierarchical Normal-Gamma prior on the regression coefficients $\boldsymbol{\beta}$ and the Normal random link between $\boldsymbol{\eta}$ and the covariates and coefficients.

Hyperparameter selection. A side effect of using the Laplace approximation for η_i is a good intuition about reasonable priors on σ^2. Such a prior should be roughly centered around the reciprocal value of the average of all counts, see (Green & Park, 2004) for details. In our implementation we use this rule of thumb by setting s_0^2 in the Inv-$\chi^2(n_0, s_0^2)$ prior on σ^2 to $1/(1 + \text{median}(\boldsymbol{z}))$. Note that the influence of the Inv-χ^2 prior can be interpreted as adding n_0 virtual samples with variance s_0^2. The only other hyperparameters in the model are the shape r and scale s in the Gamma(r, s) prior on ρ. Testing for suitable values can be done straightforwardly by first sampling ρ from the hyperprior and in turn sampling λ_g^2 from Gamma$(\frac{p_g+1}{2}, \frac{2}{p_g\rho})$. The fraction of "large" λ_g^2 values encodes our prior belief about the sparsity of the model: since λ_g^2 is a variance parameter for the g-th group of regression coefficients, the fraction of large λ_g^2 values essentially determines the expected sparsity. As a default setting in our implementation we use the criterion that roughly 1% of the λ_g^2 values shoud exceed $5 \cdot \text{median}(\boldsymbol{\lambda}^2)$.

4. Simulated Example

In order to illustrate the performance of the Bayesian Group Lasso, experiments were carried out on simulated data. The data was constructed by assuming 8 categorical variables with 3 levels each (main effects) and all higher order interaction terms upto 2nd order (total of 92 groups representing the interaction terms, and 6561 combinations of levels). The orthogonal (6561×577) design matrix X was generated with polynomial contrast codes as described in the previous section. Then three factors were chosen for generating the counts, namely one main effect (variable 1), a first order interaction (1:5) and a second order interaction (6:7:8). For these factors, positive values of β were taken, with all other β values fixed to zero. The counts were then generated using eq. (12) and eq. (13) with $\sigma^2 = 0.1$. Hyperparameters were specified using our default procedure described above. The example traceplot in the lower panel of Figure 2 indicates that the Markov chain converges almost immediately, an observation which is corroborated by a length control diagnosis according to (Raftery & Lewis, 1992) indicating that the necessary burn-in-period is probably $\ll 100$ samples.

Gibbs sampling was executed for 1000000 iterations, and the posterior distributions of the coefficients β_i were analyzed based on every 25th sample. The upper panel of Figure 2 shows the resulting estimation of significant interaction terms, with significance measured by either the fraction of positive or negative samples, depending on the sign of the mean value $\bar{\beta}_i$. The size of the circles encodes this significance measure for the main effects; the linewidth of the blue edges (1st order interactions) and reddish trian-

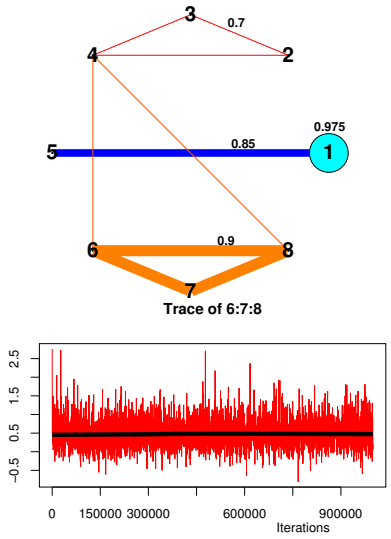

Figure 2. Simulated data: 8 categorical variables with 3 levels each, and three "truly" nonzero interaction terms: 1, 1:5, 6:7:8. **Upper panel:** interactions identified from the posterior distributions. The size of the circles indicates the estimated significance of the main effects: 97.5% of the posterior samples for variable 1 have a positive sign. Correspondingly, the linewidth of the interactions (blue lines: 1st-order, reddish triangles: 2nd-order) indicates their significance. **Lower panel:** Example traceplot and moving average for the 2nd-order interaction 6:7:8

gles represents the importance of the higher-order terms, ranging from 0.7 (thinnest lines plotted) to 0.975. Note that a level of 0.5 represents complete randomness. All three interaction terms could be clearly identified. There are 3 additional spurious interaction terms whose significance level, however, is very low (0.7).

5. Application to Breast Cancer Studies

Breast cancer and immunohistochemistry. In Western societies breast cancer is one of the leading causes of tumor-induced death in women. Despite improvements in the identification of prognostic and predictive parameters, novel biomarkers are needed to improve patient risk stratification and to optimize patient outcome. Furthermore, the identification of molecules that are differentially regulated during tumorigenesis may lead to the development of personalized therapeutic approaches.

Recently, independent research groups were able to identify five distinct gene expression patterns which are (i) highly predictive for the patients' prognosis and (ii) may reflect the biological behavior better compared to established parameters. According to this model, a basal as well as two distinct luminal-like expression patterns in addition to a **her2** (**ERBB2**) overexpressing and a normal breast-like group could be distinguished (Perou et al., 2000).

Even though results from mRNA expression profiling are very convincing, there are still some limitations to its clinical application due to the high costs. Several studies have shown that biologically distinct classes of breast cancer as defined by mRNA expression analysis can also be identified with cost efficient immunohistochemistry (Abd El-Rehim et al., 2005; Diallo-Danebrock et al., 2007). Definitions of the basal phenotype by the former and other groups using different cytokeratin antibodies (e.g. anti-**CK5/6**) prove to be robust and allow the identification of this tumor type on a routine basis.

Tissue microarrays. In the present study, intensity levels of the following immunohistochemical markers in tissue samples have been measured utilizing the *tissue microarray* (TMA) technology: the estrogen receptor (**er**), karyopherin-alpha-2 (**KPNA2**) , anti-cytokeratin **CK5/6**, fibrous structural protein **Collagen-6**, membrane-associated tetraspanin protein **Claudin-7**, inter-α-trypsin inhibitor **ITIH5**, and the human epidermal growth factor receptor **her2**. The TMA technology promises to significantly accelerate studies seeking for associations between molecular changes and clinical endpoints (Kononen & Bubendorf, L. et al, 1998).

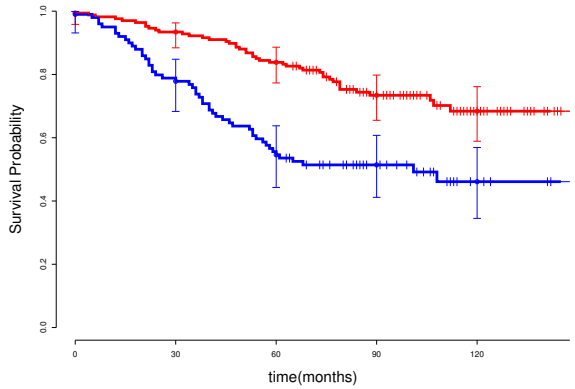

Figure 3. Kaplan-Meier curves regarding overall survival for the low-risk (upper red curve) and the high-risk group (lower blue curve) of breast cancer patients. Error bars define standard 95% confidence intervals.

Experimental design. After histopathological grading of tumors according to (Elston & Ellis, 1991), patients were divided into a low-risk (grade 1-2) and a high-risk (grade 3) group. The overall goal of this experiment was to identify differences in interaction patterns of marker proteins between these groups. Kaplan-Meier survival analysis in Figure 3 shows that the split chosen is meaningful in the sense that the survival probability differs significantly between these two groups. This observation was corroborated by the analysis of mean protein expression levels in the two groups (Figure 4). As expected for the low-risk class of patients (grade 1-2), there was marked estrogen receptor (**er**) expression, whereas **CK5/6** and **KPNA2** were negative.

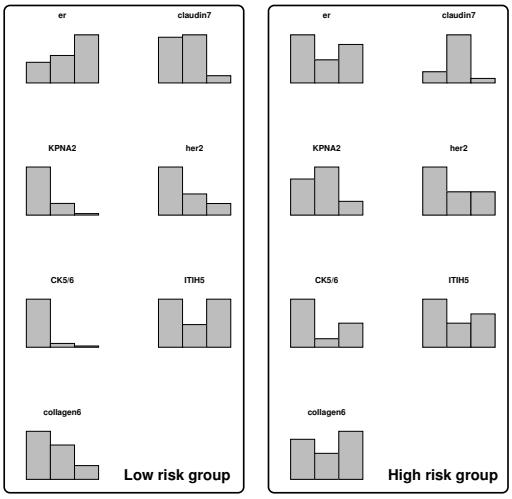

Figure 4. Distribution of protein expression levels (3 bins corresponding to "low", "intermediate", "high") in the low-risk group (left) and the high-risk group (right).

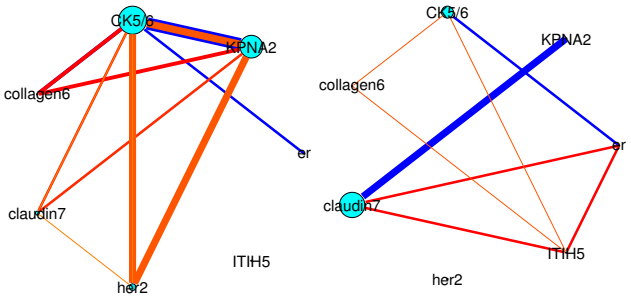

Figure 5. Identified interaction patterns for the low-risk group (left) and the high-risk group (right). The size of the circles indicates the estimated significance of the main effects. For instance, the largest circle for **CK5/6** means that more than 95% of the posterior samples are negative. Correspondingly, the linewidth of the interactions (blue lines: 1st-order, reddish triangles: 2nd-order) indicates their significance, see Figure 6 for an example.

Data analysis and interpretation. Having identified a meaningful subdivision into patient groups, we focused on identifying interaction patterns in each of the groups. The observed expression of each of the proteins was represented as a factor with 3 levels ("low", "intermediate", "high"). The resulting contingency tables were separately analyzed for each group with our Bayesian Poisson Group-Lasso model. Interaction terms up to the second order (i.e. interactions between triples of factors) were analyzed. One million Gibbs samples were drawn, the burn-in phase contained the first 200,000 samples, and every 25th of the remaining samples was used for computing the posterior densities. Due to our efficient algorithm, it took less than one hour to compute all 1 000 0000 samples on a standard computer. Traceplots indicated that the convergence of the Markov chain was not really an issue in this experiment (see Figure 6 for an example), an observation which is corroborated by a length control diagnosis according to (Raftery & Lewis, 1992) indicating that the necessary burn-in-period is probably $\ll 1000$.

In the low-risk group the following interaction terms appeared to be highly significant: the two main effects of **KPNA2** and **CK5/6** expression, the first-order interaction **KPNA2**:**CK5/6** and the second order interaction **KPNA2**:**CK5/6**:**her2**, see Figure 5. Interpreting high-order interaction terms can be a complex problem. A close analysis of the contrast codes and the sign of the regression coefficients showed, however, that all these interaction terms explain observed counts by either marginal or joint negative immunoreactivity for **KPNA2**, **CK5/6** and **her2**, respectively.

The high-risk group showed a distinctly different interaction pattern which is dominated by the main effect of **claudin7** expression, the interaction **claudin7**:**KPNA2** and the (weaker) interaction **er**:**claudin7**:**ITIH5**. Again, looking into the contrast codes and the signs, we concluded that the main effect of claudin7 explains the counts by an over-represented "intermediate" bin and both under-represented "low" and "high" bins. The interaction term **claudin7**:**KPNA2** explains counts mainly by a joint "intermediate" expression.

These interaction patterns are in line with known gene expression patterns of breast cancer. Non-high grade breast cancers (grade 1-2) were mainly hormone-receptor positive, and negative for the high-grade markers **KPNA2**, **CK5/6** and **her2**. In a study by Dahl et al. (Dahl et al., 2006), high rates of **KPNA2** expression were significantly associated with positive **TP53** and **her2** immunoreactivity and a high proliferation index. Besides **CK5/6**, **KPNA2** seemed to be characteristic of the basal-like subtype of breast cancers, possibly representing a different clinical entity of breast tumors, which is associated with shorter survival times and a high frequency of **TP53** mutations.

Control experiments. In order to compare our results with other analysis methods we conducted two control experiments. For the low-risk group, Figure 7 shows the "solution path" computed by the non-Bayesian analogue of our method, the standard Group-Lasso with Poisson likelihood. We used the algorithm described in (Roth & Fischer, 2008). The solution path shows the evolution of the individual group norms when relaxing the constraint κ, see eq. (3). The plot indicates that the main effects **CK5/6** and **KPNA2** and the interactions **KPNA2**:**CK5/6** and **KPNA2**:**CK5/6**:**her2** have a dominating role, which is in perfect agreement with our results. At the same time, the more diffuse picture for large constraint values $\kappa > 100$ to-

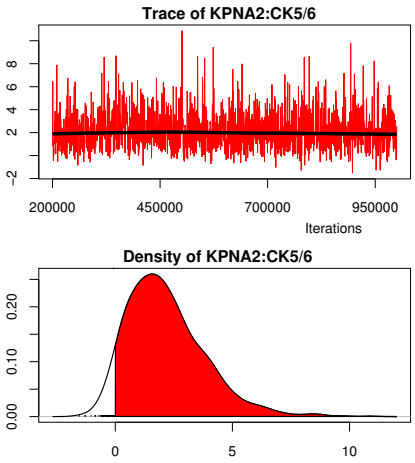

Figure 6. Example traceplot for to the very strong 1st-order-interaction **KPNA2**:**CK5/6** in the low-risk group, and moving average (upper panel). Corresponding posterior density (lower panel). More than 90% of the samples exceed zero (red area).

gether with the difficulty of defining meaningful variance estimates nicely demonstrates the inherent interpretation problems of classical Group-Lasso solutions.

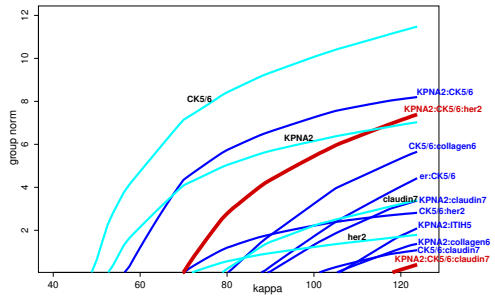

Figure 7. Comparison with non-Bayesian analogue for the low-risk group: evolution of group norms ("solution path") obtained by relaxing the constraint in the standard Group-Lasso with Poisson likelihood.

The test for uniqueness/completeness of solutions proposed in (Roth & Fischer, 2008) reveals another problem: for any reasonable numerical tolerance parameter in the optimization process, the solutions found by the Group-Lasso are probably not uniquely identifiable. For constraint values $\kappa > 90$ there is an increasing amount of inactive (i.e. zero norm) groups that might become active in alternative solutions which are ϵ-close (in terms of likelihood) to the found "optimal" solution. This problem might be viewed as another strong argument for following the Bayesian paradigm of averaging over Group-Lasso solutions, instead of focusing on a single (penalized) maximum likelihood solution.

For a second control experiment we used the same data to estimate a Bayesian network. Concerning the identification of interactions, the main technical differences to our Group-Lasso model are the restriction to a graph (instead of a hypergraph), and the use of directed edges which in some cases can be used for inferring causal relations. We used the *deal*-Package (Bøttcher & Dethlefsen, 2003; Bøttcher & Dethlefsen., 2009) that finds the topology on the basis of the *network score* which is basically the log of the joint probability of the graph and the data. In the resulting analysis, we observe many similarities between the Markov blankets from the Bayes nets and from our model. On the other hand, neither the network topology nor the direction of edges seems to be very stable. Among the top-scoring models many variants of the network have almost indistinguishable scores. Most of these fluctuation concern the dependencies between the three variables **KPNA2**, **claudin7** and **her2**, see Figure 8 for an example. The clear identification of a second-order interaction between these three variables in our model (the bold red triangle in the left panel of Figure 5) might be interpreted as a strong advantage of explicitly modeling high-order interactions in a hypergraph.

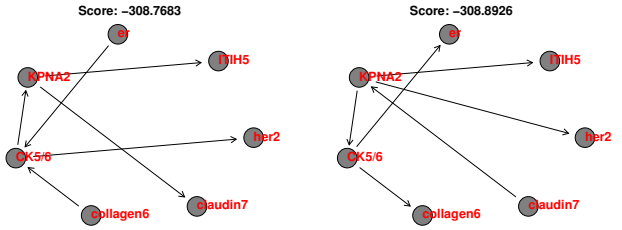

Figure 8. Two examples of the top-scoring Bayes nets with almost indistinguishable scores, showing typical variations in topology. For quantifying the score differences we made a perturbation experiment: when randomly leaving out 10% of the observations, the standard deviation of the individual maximum scores is $\sigma \approx 2.2$. More than 50 different networks in the unperturbed problem lie within the highest one-σ region.

6. Conclusion

This paper has presented a full Bayesian treatment of the Group-Lasso which is applicable to many real-world learning problems. Despite the generality of the proposed model, the main focus was on Poisson likelihood models for analyzing count data in contingency tables, with feature selection, because (i) they occur frequently in a bio-medical context, (ii) they are endowed with an inherent group structure representing the interaction terms of the categorical variables and (iii) they allow for efficient Gibbs sampling models making it practical in real-world examples.

On the theoretical side we have derived a hierarchical Normal-Gamma expansion of the Multi-Laplace prior on the regression coefficients. This expansion is one of the key components for rewriting the Group-Lasso model in a form that is suitable for efficient Gibbs sampling. The second key component is the orthogonality property of the design matrix representing dummy encodings for categori-

cal variables which ensures that sampling becomes numerically stable and highly efficient.

When applied to a real-world breast-cancer study, the proposed method identifies the differences in interactions patterns of marker proteins between multiple patient groups. To our knowledge, this is the first study which systematically analyzes the influence of high-order interactions based on immunohistochemical data for breast cancer. Interpretation of the results by pathologists further validates our approach since the findings are in line with known gene expression patterns of breast cancer and hence supporting the usage of cost-effective immunohistochemical markers.

The comparison of the method to standard approaches (standard Group-Lasso, Bayesian network) validates the results and also highlights the advantages of the proposed method over existing approaches. The related software is going to be published as an R package.

Acknowledgments

The work was partly financed with a grant of the Swiss SystemsX.ch Initiative to the project "LiverX" of the Competence Center for Systems Physiology and Metabolic Diseases. The LiverX project was evaluated by the Swiss National Science Foundation.

References

Abd El-Rehim, D., Ball, G., & Pinder, S.E. et al. (2005). High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *Int J Cancer*, *116*, 340–50.

Bøttcher, S. G., & Dethlefsen, C. (2003). deal: A package for learning bayesian networks. *Journal of Statistical Software*, *8*, 200–3.

Bøttcher, S. G., & Dethlefsen., C. (2009). *deal: Learning bayesian networks with mixed variables*. R package version 1.2-33.

Dahinden, C., Parmigiani, G., Emerick, M., & Bühlmann, P. (2007). Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. *BMC Bioinformatics*, *8*, 476.

Dahl, E., G., G. K., & Gottlob, K. et al. (2006). Molecular profiling of laser-microdissected matched tumor and normal breast tissue identifies karyopherin alpha2 as a potential novel prognostic marker in breast cancer. *Clin Cancer Res*, *12*, 3950–60.

Diallo-Danebrock, R., Ting, E., & Gluz, O. et al. (2007). Protein expression profiling in high-risk breast cancer patients treated with high-dose or conventional dose-dense chemotherapy. *Clin Cancer Res*, *13*, 488–97.

Elston, C., & Ellis, I. (1991). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, *19*, 403–410.

Figueiredo, M., & Jain, A. (2001). Bayesian learning of sparse classifiers. *Proc. IEEE Comp. Soc. Conf. Computer Vision and Pattern Recognition* (pp. 35–41).

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (1995). *Bayesian data analysis*. Chapman&Hall.

Green, P., & Park, T. (2004). Bayesian methods for contingency tables using Gibbs sampling. *Statistical Papers*, *45*, 33–50.

Kim, Y., Kim, J., & Kim, Y. (2006). Blockwise sparse regression. *Statistica Sinica*, *16*, 375–390.

Kononen, J., & Bubendorf, L. et al (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine*, *Jul;4(7)*, 844–7.

Meier, L., van de Geer, S., & Bühlmann, P. (2008). The Group Lasso for Logistic Regression. *J. Roy. Stat. Soc. B*, *70*, 53–71.

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*, 681–686.

Perou, C., Sorlie, T., & Eisen, M.B. et al. (2000). Molecular portraits of human breast tumours. *Nature*, *406*, 747–752.

Raftery, A., & Lewis, S. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, *7*, 493–497.

Roth, V., & Fischer, B. (2008). The Group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. *ICML '08: Proceedings of the 25th international conference on Machine learning* (pp. 848–855). New York, NY, USA: ACM.

Seeger, M. (2008). Bayesian inference and optimal design in the sparse linear model. *Journal of Machine Learning Research*, *9*, 759–813.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc. B*, *58*, 267–288.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. B*, 49–67.