Journal of Machine Learning Research 17 (2016) 1-35 Submitted 10/14; Revised 4/16; Published 6/16

Multiple Output Regression with Latent Noise

Jussi Gillberg JUSSI.GILLBERG@AALTO.FI
Pekka Marttinen PEKKA.MARTTINEN@AALTO.FI
Helsinki Institute for Information Technology HIIT
Department of Computer Science
PO Box 15600, Aalto University, 00076 Aalto, Finland

Matti Pirinen MATTI.PIRINEN@HELSINKI.FI
Institute for Molecular Medicine Finland (FIMM)
University of Helsinki, Finland

Antti J. Kangas ANTTI.KANGAS@COMPUTATIONALMEDICINE.FI
Pasi Soininen * PASI.SOININEN@COMPUTATIONALMEDICINE.FI
Computational Medicine
Faculty of Medicine
University of Oulu & Biocenter Oulu, Oulu, Finland

Mehreen Ali MEHREEN.ALI@HELSINKI.FI
Institute for Molecular Medicine Finland (FIMM)
University of Helsinki, Finland

Aki S. Havulinna AKI.HAVULINNA@THL.FI
Department of Health
National Institute for Health and Welfare, Helsinki, Finland

Marjo-Riitta Järvelin* M.JARVELIN@IMPERIAL.AC.UK
Department of Epidemiology and Biostatistics
MRC-PHE Centre for Environment & Health, School of Public Health,
Imperial College London, UK

Mika Ala-Korpela* MIKA.ALA-KORPELA@COMPUTATIONALMEDICINE.FI
Computational Medicine
Faculty of Medicine
University of Oulu & Biocenter Oulu, Oulu, Finland

Samuel Kaski SAMUEL.KASKI@AALTO.FI
Helsinki Institute for Information Technology HIIT
Department of Computer Science
PO Box 15600, Aalto University, 00076 Aalto, Finland

Editor: Karsten Borgwardt

∗. PS and MAK are also at NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland; MRJ is also at Center for Life Course Epidemiology, Faculty of Medicine, University of Oulu, Finland and Biocenter Oulu, University of Oulu, Finland and Unit of Primary Care, Oulu University Hospital, Oulu, Finland; MAK is also at Computational Medicine, School of Social and Community Medicine and the Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK

Abstract

In high-dimensional data, structured noise caused by observed and unobserved factors affecting multiple target variables simultaneously, imposes a serious challenge for modeling, by masking the often weak signal. Therefore, (1) explaining away the structured noise in multiple-output regression is of paramount importance. Additionally, (2) assumptions about the correlation structure of the regression weights are needed. We note that both can be formulated in a natural way in a latent variable model, in which both the interesting signal and the noise are mediated through the same latent factors. Under this assumption, the signal model then borrows strength from the noise model by encouraging similar effects on correlated targets. We introduce a hyperparameter for the *latent signal-to-noise ratio* which turns out to be important for modelling weak signals, and an ordered infinite-dimensional shrinkage prior that resolves the rotational unidentifiability in reduced-rank regression models. Simulations and prediction experiments with metabolite, gene expression, FMRI measurement, and macroeconomic time series data show that our model equals or exceeds the state-of-the-art performance and, in particular, outperforms the standard approach of assuming independent noise and signal models.

Keywords: Bayesian reduced-rank regression, latent variable models, latent signal-to-noise ratio, multiple-output regression, nonparametric Bayes, shrinkage priors, structured noise, weak effects

1. Introduction

Explaining away structured noise is one of the cornerstones for successful modeling of high-dimensional output data in the regression framework (Fusi et al., 2012; Klami et al., 2013; Rai et al., 2012; Rakitsch et al., 2013; Stegle et al., 2012; Virtanen et al., 2011). The structured noise refers to dependencies between response variables, which are unrelated to the dependencies of interest between the response variables and the covariates. It is noise caused by observed and unobserved confounders that affect multiple variables simultaneously. Common observed confounders in medical and biological data include age and sex of an individual, whereas unobserved confounders include, for example, the state of the cell being measured, measurement artifacts influencing multiple probes, or other unrecorded experimental conditions. When not accounted for, structured noise may both hide interesting relationships and result in spurious findings (Leek and Storey, 2007; Kang et al., 2008).

The effects of known confounders can be removed straightforwardly by using supervised methods. For the unobserved confounders, a routinely used approach for explaining away structured noise has been to assume *a priori* independent effects for the interesting and uninteresting factors. For example, in the factor regression setup (West, 2003; Stegle et al., 2010; Fusi et al., 2012), the target variables Y are assumed to have been generated as

$$Y = X\mathbf{\Theta} + H\Lambda + E, \tag{1}$$

where $Y_{N\times K}$ is the matrix of K target variables (or dependent variables) and $X_{N\times P}$ contains the covariates (or independent variables), for the N observations. The model parameter matrix $H_{N\times S_2}$ comprises the unknown latent factors and $\Lambda_{S_2\times K}$ the factor loadings, which are used to model away the structured noise. The term $E_{N\times K}$ represents independent unstructured noise and the elements of E are independently distributed, $\mathrm{vec}(E) \sim \mathcal{N}(0, I_{NK})$. In this paper we call this model **independent-noise BRRR**. To reduce the effective number of parameters in the regression coefficient matrix $\mathbf{\Theta}_{P\times K}$, a low-rank structure may be

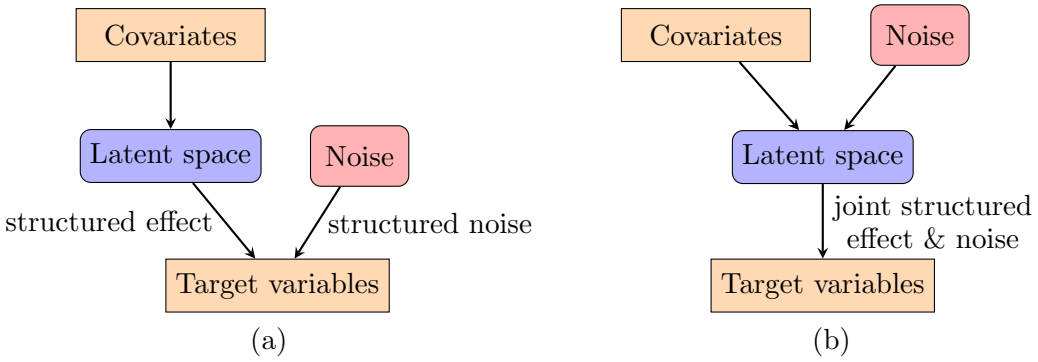


Figure 1: Illustration of (a) *a priori* independent interesting and uninteresting effects and (b) the latent noise assumption. Latent noise is mediated to the target variable measurements through a common subspace with the interesting effects.

assumed:

$$\mathbf{\Theta} = \Psi\,\Gamma, \tag{2}$$

where the rank S_1 of parameters $\Psi_{P\times S_1}$ and $\Gamma_{S_1\times K}$ is substantially lower than the number of target variables K and covariates P. The low-rank decomposition of the regression coefficient matrix (2) may be given an interpretation whereby the covariates X affect S_1 latent components with coefficients specified in Ψ, and the components, in turn, affect the target Y with coefficients Γ. Another line of work in multiple output prediction has focused on borrowing information from the correlation structure of the target variables when learning the regression model. The intuition stems from the observation that correlated targets are often seen to be affected similarly by the covariates, for example in genetic applications (see, e.g., Davis et al., 2014; Inouye et al., 2012). One popular method, GFlasso (Kim et al., 2009), learns the regression coefficients using

$$\begin{aligned}\hat{\Theta} = \operatorname{argmin} &\sum_k (\mathbf{y_k} - X\theta_\mathbf{k})^T(\mathbf{y_k} - X\theta_\mathbf{k}) + \\ &\lambda \sum_j \sum_k |\theta_{jk}| + \gamma \sum_{(m,l)\in E} r_{ml}^2 \sum_j |\theta_{jm} - \operatorname{sign}(r_{ml})\theta_{jl}|,\end{aligned} \tag{3}$$

where the $\theta_\mathbf{k}$ are the columns of $\hat{\Theta}$. Two regularization parameters are introduced: λ represents the standard Lasso penalty, and γ encourages the effects θ_{jm} and θ_{jl} of the jth covariate on correlated outputs m and l to be similar. Here r_{ml} represents the correlation between the mth and lth phenotypes. The E is an *a priori* specified correlation graph for the output variables, with edges representing correlations to be accounted for in the model.

In this paper we propose a model that simultaneously learns the structured noise and encourages the sharing of information between the noise and the regression models. To motivate the new model, we note that by assuming independent prior distributions on Γ and Λ in model (1), one implicitly assumes independence of the interesting and uninteresting effects, caused by covariates X and unknown factors H, respectively (Fig. 1a). The assumption is appealing for example when explaining away batch effects (Fusi et al., 2012)

in high-dimensional data, but may be inadequate in the presence of other types of noise in molecular biology, where gene expression and metabolomics measurements record concentrations of compounds generated by ongoing latent biological processes. In this kind of situations, a limited set of covariates, such as single nucleotide polymorphisms (SNPs), determines the activity of the latent process only partially and all other activity of the process is due to unrecorded factors. In such cases, the noise affects the measurement levels through the very same process as the interesting signal (Fig. 1b), and rather than assuming independence of the effects, an assumption about parallel effects would be more appropriate. We refer to this type of noise as *latent noise* as it can be considered to affect the same latent subspace as the interesting effects. We note that in practice both types of structured noise are likely to be present. In this work, our main focus is on the latent noise, but we also present a comparison with a model that includes both types of structured noise simultaneously.

A natural way to encode the assumption of latent noise is to use the following model structure:

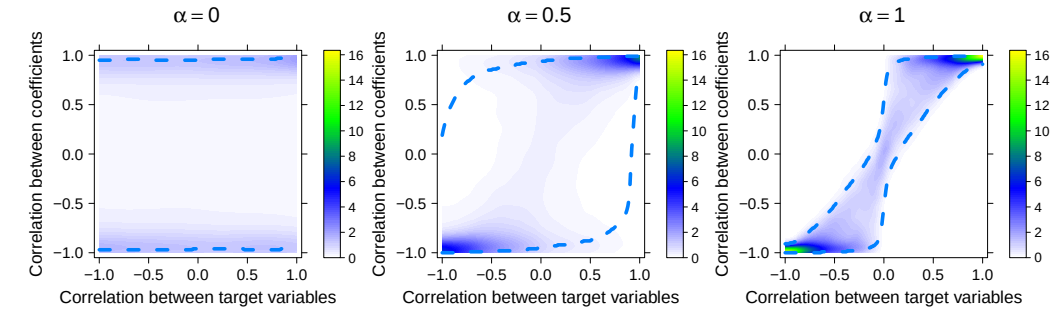
$$Y = (X\Psi + \Omega)\,\Gamma + E, \tag{4}$$

where the $\Omega_{N\times S_1}$ is a matrix consisting of unknown latent factors. In (4), Γ mediates the effects of both the interesting and uninteresting signals on the target variables. We note that the change required in the model structure is small, and has in fact been presented earlier (Bo and Sminchisescu 2009; recently extended with an Indian Buffet Process prior on the latent space by Bargi et al. 2014). We now proceed to using the structure (4) for GFlasso-type sharing of information (3) between the regression and noise models while simultaneously explaining away structured noise. To see that the information sharing between noise and regression models follows immediately from model (4), one can consider simulations generated from the model. The *a priori* independence assumption of model (1) results in uncorrelated regression weights regardless of the correlations between target variables (Figure 2a). The assumption of latent noise (4), however, encourages the regression weights to be correlated in a similar way as the target variables are (Figure 2c).

In this work, we focus on modelling weak signals in high-dimensional data with structured noise, where we consider effects that explain a tiny portion, say $< 1\%$, of the variance of the target variables as weak. We have hypothesized above that a model with the structure (4) might be particularly well-suited for this purpose. Additionally, (i) particular emphasis must be put on defining adequate prior distributions to distinguish the weak effects from noise as effectively as possible, and (ii) scalability to large sample size is needed in order to have any chance of learning the weak effects. For (i), we define *latent signal-to-noise ratio* β as a generalization of the standard signal-to-noise ratio in the latent space:

$$\beta = \frac{\mathrm{Trace}(\mathrm{Var}(X\ \Psi))}{\mathrm{Trace}(\mathrm{Var}(\Omega))}, \tag{5}$$

We use the latent signal-to-noise ratio as a hyperparameter in our model, and show that it is a key parameter affecting model performance. It can be either learned or set using prior knowledge. In addition, we introduce an ordered infinite-dimensional shrinkage prior that resolves the inherent rotational ambiguity in the model (4), by sorting both signal and noise components by their importance. Finally, we present efficient inference methods for the model.



(a) (b) (c)

Figure 2: Conditional distribution of the correlation between regression coefficients, given the correlation between the corresponding target variables. In (a) the model (1) assumes *a priori* independent regression and noise models, and in (c) the model (4) makes the latent noise assumption. (b) A mixture of the models in a and c. The data were generated using equation (18), as described in Section 5.3, and α denotes the relative proportion of latent noise in data generation. The dashed lines denote the 95% confidence intervals of the conditional distributions.

2. Related work

Simultaneously solving multiple real-valued prediction tasks with the same set of covariates is called multiple-output regression (Breiman and Friedman, 1997); and more generally sharing of statistical strength between related tasks is called multitask learning (Baxter, 1996; Caruana, 1997). The data consist of N input-output pairs $(\mathbf{x}_n, \mathbf{y}_n)_{n=1,\ldots,N}$; the P-dimensional input vectors $\mathbf{x}$ (covariates) are used for predicting K-dimensional vectors $\mathbf{y}$ of target variables. The common approach to dealing with structured noise due to unobserved confounders is to apply factor regression modeling (1) (West, 2003) and to explain away the structured noise using a noise model that is assumed to be *a priori* independent of the regression model (Stegle et al., 2010; Fusi et al., 2012; Rai et al., 2012; Virtanen et al., 2011; Klami et al., 2013; Rakitsch et al., 2013). A recent Bayesian reduced-rank regression (BRRR) model (Marttinen et al., 2014) implements the routine assumption of the independence of the regression and noise models; we will include it in the comparison studies of this paper.

Methods for multiple-output regression without the structured noise model have been proposed in other fields. In the application fields of genomic selection and multi-trait quantitative trait loci mapping, solutions (Yi and Banerjee, 2009; Xu et al., 2009; Calus and Veerkamp, 2011; Stephens, 2013) for low-dimensional target variable vectors ($K < 10$) have been proposed, but these methods do not scale up to the currently emerging needs of analyzing higher-dimensional target variable data. Additionally, sparse multiple-output regression models have been proposed for prediction of phenotypes from genomic data (Kim et al., 2009; Sohn and Kim, 2012).

Many methods for multi-task learning have been proposed in the field of kernel methods (Evgeniou and Pontil, 2007). These methods do not, however, scale up to data sets with

several thousands of samples, required for predicting the weak effects. Other relevant work include a recent method based on the BRRR presented by Foygel et al. (2012), but it does not scale to the dimensionalities of our experiments either. Methods for high-dimensional phenotypes have been proposed in the field of expression quantitative trait loci mapping (Bottolo et al., 2011) for the related task of finding associations (and avoiding false positives) rather than prediction, which is our main focus. Also functional assumptions (Wang et al., 2012) have been used to constrain related learning problems.

3. Model

In this Section, we present the details of our new model, Bayesian reduced rank regression with latent noise (latent-noise BRRR), show how the hyperparameters can be set using the latent signal-to-noise ratio, and analyze theoretically some properties of the infinite-dimensional shrinkage prior.

3.1 Model details: latent-noise BRRR

Our model is given by

$$Y = (X\Psi + \Omega)\Gamma + E, \tag{6}$$

where $Y_{N\times K}$ contains the K-dimensional response variables for N observations, and $X_{N\times P}$ contains the predictor variables. The product $\Theta = \Psi\Gamma$, of $\Psi_{P\times S_1}$ and $\Gamma_{S_1\times K}$, results in a regression coefficient matrix with rank S_1. The $\Omega_{N\times S_1}$ contains unknown latent factors representing the latent noise. Finally, $E_{N\times K} = [e_1, \ldots, e_N]^T$, with $e_i \sim N(0, \Sigma)$, where $\Sigma = diag(\sigma_1^2, \ldots, \sigma_K^2)$ is a matrix of uncorrelated target variable-specific noise vectors. Figure 3 displays graphically the structure of the model. In the figure, the node corresponding to the parameter Γ that is shared by the regression and noise models is highlighted with green.

Similarly to a recent BRRR model (Marttinen et al., 2014) and the Bayesian infinite sparse factor analysis model (Bhattacharya and Dunson, 2011), we assume the number of components S_1 connecting the covariates to the targets to be infinite. Accordingly, the number of rows in the weight matrix Γ, and the numbers of columns in Ψ and Ω, are infinite. The low-rank nature of the model is enforced by shrinking the columns of Ψ and rows of Γ and Ω increasingly with the growing column/row index, such that only a small number of columns/rows are influential in practice. The increasing shrinkage also solves any rotational unidentifiability issues by enforcing the model to mediate the strongest effects through the first columns/rows. In Section 3.4 we explore the basic properties of the infinite-dimensional prior, to ensure its soundness. The hierarchical priors for the projection weight matrix Γ, where $\Gamma = [\gamma_{hj}]$, are set as follows:

$$\begin{aligned} \gamma_{hj}|\phi_{hj}^{\Gamma}, \tau_h &\sim N\left(0, \left(\phi_{hj}^{\Gamma}\tau_h\right)^{-1}\right), \ \phi_{hj}^{\Gamma} \sim \mathrm{Ga}(\nu/2, \nu/2), \\ \tau_h = \prod_{l=1}^{h} \delta_l, &\quad \delta_1 \sim \mathrm{Ga}(a_1, 1), \quad \delta_l \sim \mathrm{Ga}(a_2, 1), \quad l \geq 2. \end{aligned} \tag{7}$$

Here τ_h is a global shrinkage parameter for the hth row of Γ and the ϕ_{hj}^{Γ}s are local shrinkage parameters for the individual elements of Γ, to provide additional flexibility over the global

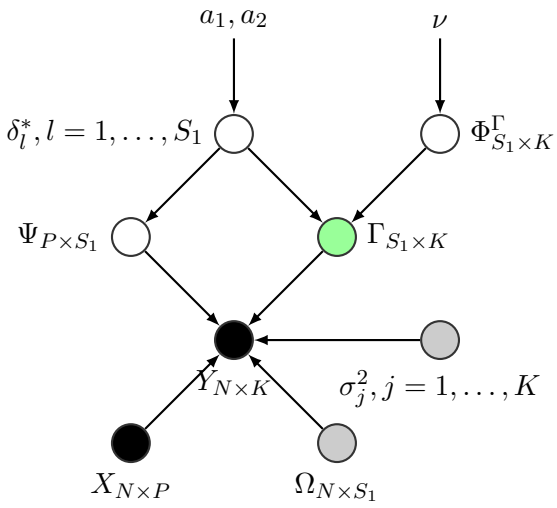


Figure 3: Graphical representation of latent-noise BRRR. The observed data are denoted by black circles, variables related to the reduced-rank regression part of the model by white circles, variables related only to the noise model are denoted by gray circles, and variables related to both the regression and the structured noise model are denoted with green circles. The matrix $\Phi^{\Gamma}_{S_1 \times K}$ comprises the sparsity parameters for the K target variables for the components.

shrinkage priors. The same parameters τ_h are used to shrink the columns of the matrices $\Psi = [\psi_{jh}]$ and $\Omega = [\omega_{jh}]$, because the scales of Γ and Ψ (or Ω) are not identifiable separately:

$$\psi_{jh}|\tau_h \sim N\left(0, (\tau_h)^{-1}\right), \quad \text{and} \quad \omega_{jh}|\tau_h \sim N\left(0, \sigma_\Omega^2 (\tau_h)^{-1}\right),$$

where σ_Ω^2 is a parameter that specifies the amount of latent noise, which is used to regularize the model (see the next Section). With the priors specified, the hidden factors Ω can be integrated out analytically, yielding

$$y_i \sim N\left((\Psi\Gamma)^T x_i, \sigma_\Omega^2 (\Gamma^*)^T (\Gamma^*) + \Sigma\right), \quad i = 1, \ldots, N, \tag{8}$$

where Γ^* is obtained from Γ by multiplying the rows of Γ with the shrinkages $(\tau_h)^{-1/2}$ of the columns of Ω.

Finally, conjugate prior distributions

$$\sigma_j^{-2} \sim \mathrm{Ga}(a_\sigma, b_\sigma), \quad j = 1, \ldots, K, \tag{9}$$

are placed on the noise parameters of the target variables.

3.2 Regularization of latent-noise BRRR through the variance of Ω

The latent signal-to-noise ratio β in Equation (5) has an intuitive interpretation: given our prior distributions for Ψ and Ω, the prior latent SNR indicates the extent to which we believe

the noise to explain variation in Y, as compared to the variance explained by the covariates X. Thus, the latent SNR acts as a regularization parameter: when the latent variables Ω are allowed to have a large variance, the data will be explained by the noise model rather than the covariates. We note that this approach to regularization is non-standard and it may have favourable characteristics compared to the commonly used L1/L2 regularization of regression weights. First of all, the regression weights remain relatively unbiased as they need not be enforced to zero to control for overfitting. This is important when the effects are weak: if the effects were shrunk towards zero, they might be lost completely.

Secondly, while regularizing with the *a priori* selected latent SNR, the regularization parameter itself remains interpretable: every value of the variance parameter of Ω can be immediately interpreted as the percentage of variance explained by the noise model as compared to the covariates. In our experiments, we use cross-validation to select the variance of Ω and the interpretability of the parameter makes it easy to express beliefs of the plausible values based on prior knowledge. Making similar educated guesses for L1/L2 regularization parameters is not straightforward.

3.3 Difference between latent-noise BRRR and independent-noise BRRR

We call the standard Bayesian reduced rank regression (Equation 1), which assumes independent noise and signal models, the *independent-noise BRRR*. The new latent-noise BRRR differs from it in two ways: in the latent-noise BRRR

1. the structure of the model is different in that the noise model uses the same projection parameters as the regression model, and

2. the model is regularized by modifying the variance of the noise model. This is achieved by learning the latent signal-to-noise ratio parameter β.

In Section 5.5 we show that both of these improvements are needed to reach the performance differences observed.

We emphasize that although the technical difference between the two models is minor, the models are very different from the conceptual point of view, as discussed in the Introduction, as well as from the practical point of view. In particular, it has been reported before that with weak effects the independent-noise BRRR may suffer from severe instability, resulting from a highly multi-modal posterior distribution and, consequently, poor convergence and mixing properties of the learning algorithms (Koop et al., 2006; Marttinen et al., 2014). In Section 5.11, we demonstrate how the latent noise assumption provides just the required additional regularization to make the formal Bayesian inference tractable even with weak effects.

As both independent structured noise and latent noise could be present, a logical extension to the models presented so far is to consider both noise types simultaneously,

$$Y = (X\Psi + \Omega)\,\Gamma + H\Lambda + E, \tag{10}$$

where the distributional assumptions for Ψ, Ω and Γ are the same as in latent-noise BRRR, and for H and Λ they follow independent-noise BRRR. The Gibbs updates for this model are straightforward modifications of those for the latent-noise BRRR and independent-noise BRRR. We have implemented also this model and study its performance in Section 5.9.

We note that the latent-noise model is, in principle, able to express data generated by the independent-noise BRRR model, and vice versa. The latent-noise BRRR model may learn noise components that are independent from the signal in practice, having negligible contribution from the regression part $X\Psi$. On the other hand, nothing prevents the independent noise model to learn some correlated regression and noise components. Therefore, the family of models defined by Equation (10) that simultaneously includes both kinds of structured noise may have redundancy in its parameters. Indeed, the experiments in Section 5.9 demonstrate only minor improvements from this model.

3.4 Proofs of the soundness of the infinite prior

In this Section we verify the sensibility of the infinite non-parametric prior, which we introduce for ordering the components according to decreasing importance, and of a computational approximation resulting from truncation of the infinite model.

It has been proven that in Bayesian factor models $a_1 > 2$ and $a_2 > 3$ (in our case defined in eqn 7) is sufficient for the elements of $\Lambda\Lambda^T$ to have finite variance in a Bayesian factor model (1), even if an infinite number of columns with a prior similar to our model is assumed for Λ (Bhattacharya and Dunson, 2011). In this Section we present similar characteristics for the infinite reduced-rank regression model. The detailed proofs can be found in the Supplementary material. First, in analogy to the infinite Bayesian factor analysis model, we show that

$$a_1 > 2 \quad \text{and} \quad a_2 > 3 \tag{11}$$

is sufficient for the prediction of any of the response variables to have finite variance under the prior distribution (Proposition 1). Second, we show that the underestimation of uncertainty (variance) resulting from using a finite rank approximation to the infinite reduced-rank regression model decays exponentially with the rank of the approximation (Proposition 2). For notational clarity, let Ψ_h denote the h^{th} column of the Ψ matrix in the following. With this notation, the prediction for the ith response variable can be written as

$$\begin{aligned}\widetilde{y_i} &= x^T\Theta_i \\ &= x^T \sum_{h=1}^{\infty} \Psi_h \gamma_{hi}.\end{aligned}$$

Furthermore, let $\Gamma(\cdot)$ denote below the gamma function (not to be confused with the matrix Γ used in all other Sections of this paper).

Proposition 1: Finite variance of predictions Suppose that $a_1 > 2$ and $a_2 > 3$. Then

$$\mathrm{Var}(\widetilde{y_i}) = \frac{\nu}{\nu - 2} \sum_{j=1}^{P} \mathrm{Var}(x_j) \frac{\Gamma(a_1 - 2)/\Gamma(a_1)}{1 - \Gamma(a_2 - 2)/\Gamma(a_2)}. \tag{12}$$

A detailed proof is provided in the Supplementary material.

Proposition 2: Truncation error of the finite rank approximation Let $\widetilde{y_i}^{S_1}$ denote the prediction for the ith target variable when using an approximation for Ψ and Γ consisting of the first S_1 columns or rows only, respectively. Then,

$$\frac{\mathrm{Var}(\widetilde{y_i}) - \mathrm{Var}(\widetilde{y_i}^{S_1})}{\mathrm{Var}(\widetilde{y_i})} = \left[\frac{\Gamma(a_2 - 2)}{\Gamma(a_2)}\right]^{S_1},$$

that is, the reduction in the variance of the prediction resulting from using the approximation, relative to the infinite model, decays exponentially with the rank of the approximation. A detailed proof is provided in the Supplementary material.

4. Efficient computation by reparameterization

For estimating the parameters of the latent-noise BRRR, we use Gibbs sampling, updating the parameters one by one by sampling them from their conditional posterior probability distributions, given the current values of all other parameters. The bottleneck of the computation is in updating the matrix Ψ, and below we present a novel efficient update for this parameter.

4.1 Update of Γ

The conditional distribution of the parameter matrix Γ of latent-noise BRRR can be updated using a standard result for Bayesian linear models (Bishop et al., 2006) which states that if

$$\beta \sim N(0, \Sigma_\beta), \quad \text{and} \quad y|X^*, \beta \sim N(X^*\beta, \Sigma_y), \tag{13}$$

then

$$\beta|y, X^* \sim N(\Sigma_{\beta|Y}(X^{*T}\Sigma_y^{-1}y), \Sigma_{\beta|y}), \tag{14}$$

where

$$\Sigma_{\beta|y} = (\Sigma_\beta^{-1} + X^{*T}\Sigma_y^{-1}X^*)^{-1}. \tag{15}$$

Because in our model (6) the columns E_i of the noise matrix are assumed independent with variances $\sigma_1^2, \ldots, \sigma_K^2$, we get

$$Y_i \sim N((X\Psi + \Omega)\Gamma_i, \sigma_i^2 I_N). \tag{16}$$

Thus, by substituting

$$X^* \leftarrow X\Psi + \Omega, \quad \beta \leftarrow \Gamma_i, \quad \text{and} \quad \Sigma_y \leftarrow \sigma_i^2 I_N$$

into (13), together with prior covariance Σ_β derived from (7), we immediately obtain the posterior of Γ_i from (14) and (15).

4.2 Updates of $\Phi^\Gamma, \delta, \sigma$ and Ω

The updates of the hyperparameters are the same as in Bayesian Reduced Rank Regression, and the conditional posterior distributions of the hyperparameters can be found in the Supplementary material of Marttinen et al. (2014). The Ω has the same conditional posterior distribution as the model parameter H of Marttinen et al. (2014).

4.3 Improved update of Ψ

The computational bottleneck of the naïve Gibbs sampler is the update of parameter Ψ, which has PS_1 elements with a joint multivariate Gaussian distribution, conditionally on the other parameters (Geweke, 1996; Marttinen et al., 2014). Thus, the inversion of the

precision matrix of the joint distribution has a computational cost of $O(P^3 S_1^3)$. To remove the bottleneck, we reparameterize our model, after which a linear algebra trick by Stegle et al. (2011) can be used to reduce the computational cost of the bottleneck to $O(P^3 + S_1^3)$. When sampling Ψ we also integrate over the distribution of Ω following the standard result from Equation (8). The reparameterization and the new posteriors are presented in the Supplementary material.

In brief, the trick is that the eigenvalue decomposition of a matrix of the form

$$C \otimes R + \sigma I \tag{17}$$

can be evaluated inexpensively. After reparameterizing the model in the proposed way the posterior covariance matrix of Ψ becomes of the form (17) and the eigenvalue decomposition can then be used to efficiently generate samples from the posterior distribution of Ψ. We note that the trick can also be applied to the original formulation of the Bayesian reduced-rank regression model by Geweke (1996) and the R-code published with this article allows generating samples from the original model as well. In the next Section, we compare the computational cost of the algorithm using the naïve Gibbs sampler and the improved version that uses the new parameterization.

4.4 Sampling the maximum rank of the model

The sparse infinite factor analysis model presented by Bhattacharya and Dunson (2011) uses a certain adaption procedure to update the maximum rank, i.e., the truncation point of their infinite-rank factor model. The idea is to update the maximum rank occasionally during the algorithm such that ranks having all elements of the corresponding projection vectors within some pre-specified distance from zero are removed from the model and, if none of the ranks has all elements within the threshold, another rank is added into the model. We have implemented a modification of this approach where we adapt the maximum rank of our infinite reduced rank regression model using a pre-specified cutoff for the amount of variance explained by the corresponding rank. With a slight abuse of terminology, we shall call this updating of the rank as sampling in the sequel.

5. Experiments

We start with a basic validation of the latent-noise BRRR model, and its relative merits over alternatives in a prediction task, using simulations with the ground truth available (Section 5.3), and a real-world omics dataset (Section 5.4). Section 5.5 analyses these results in more detail and identifies the characteristics of the proposed latent-noise BRRR model that are responsible for the performance differences observed, by considering the impact of each novel model aspect in isolation. Section (5.7) investigates another application domain, the detection of multivariate associations. In order to assess the prediction performance in more general, we analyse several additional real-world data sets from different domains in Section 5.8.

Different aspects of the inference algorithm are considered in three sub-sections: sampling vs. cross-validation of the rank and the latent signal-to-noise ratio (Section 5.6), speedup resulting from the proposed re-parameterization of the algorithm (Section 5.10),

and convergence diagnostics (Section 5.11). To assess the value of further extensions, Section 5.9 considers a model that includes both latent and independent structured noise simultaneously. Finally, Section 5.12 summarizes the findings on all real data sets.

5.1 Data sets

Experiments were performed on the following data sets:

NFBC1966 [$N = 4702, P = 101, K = 96$, metabolomics prediction from SNPs] The NFBC1966 data set comprises genome-wide SNP data along with metabolomics measurements for a cohort of 4,702 individuals (Rantakallio, 1969; Soininen et al., 2009). With these data, 96 metabolites belonging to the subclasses VLDL, IDL, LDL and HDL (Inouye et al., 2012) were used as the target variables and SNPs known to be associated with lipid metabolism (Teslovich et al., 2010; Kettunen et al., 2012; Global Lipids Genetics Consortium, 2013) were used as the covariates. Effects of age, sex, and lipid lowering medication were regressed out from the metabolomics data as a preprocessing step. For the genotype data, SNPs with low minor allele frequency (<0.01) were removed as a preprocessing step. For this data set, the comparison method GFlasso required excessive training time and we used 5-fold cross-validation to evaluate test set performances. Where cross-validation was needed for selecting model parameter values, the validation data performance was measured as an average over 3 validation sets, each comprising $\frac{1}{10}$ of the training samples.

DILGOM [$N = 509, P = 65, K = 18 \ldots 137$, metabolomics and gene expression prediction from SNPs] The DILGOM data set (Inouye et al., 2010) consists of genome-wide SNP data along with metabolomics and gene expression measurements. For details concerning metabolomics and gene expression data collection, see Soininen et al. (2009) and Kettunen et al. (2012). In total 509 individuals had all three measurement types. The DILGOM metabolomics data comprises 137 metabolites, most of which represent NMR-quantified levels of lipoproteins classified into 4 subclasses (VLDL, IDL, LDL, HDL), together with quantified levels of amino acids, some serum extracts, and a set of quantities derived as ratios of the aforementioned metabolites. All 137 metabolites were used simultaneously as prediction targets. In gene expression prediction, in total 387 probes corresponding to curated gene sets of 8 KEGG lipid metabolism pathways were used as the prediction targets. A separate model was learnt for each pathway. The average number of probes in a pathway was 48. For details about the pathways, see the Supplementary material. On these data sets, 10-fold cross-validation was used to evaluate test set performances. To select values of the parameters that required evaluation on validation data, the training data was then further divided into 9 folds, on which cross-validation was performed to select parameters according to averaged validation set performance.

fMRI [$N = 1307, P = 776, K = 250$, fMRI response prediction from text stimuli] The cognitive neuroscience data set (Wehbe et al., 2014) consists of a time series of fMRI measurements from 8 subjects reading a chapter from "Harry Potter and the Sorcerers Stone" using *Rapid Serial Visual Presentation*: words of the text are presented one by one in the center of a screen. Brain voxel activations were measured

every 2 seconds. The 250 most accurately predictable voxels (see Supplementary material of Wehbe et al., 2014) of the fMRI measurements were used as prediction targets. The fMRI measurements from all patients were predicted simultaneously from features of the words being shown, such as semantic and syntactic properties, visual properties and discourse level features. The data were divided into 10 folds, only two of which were used to measure test data performance. This computational compromise was needed as the preprocessing (Wehbe et al., 2014) for each fold required about 10,000 hours of computation. To select the values of parameters that required evaluation on validation data, the training data were further divided into 10 folds, on which cross-validation was performed to select parameters according to averaged validation set performance.

econ [$N = 120, P = 52, K = 52,$ macroeconomic time series prediction] The macroeconomic time series data set (Stock and Watson, 2006) consists of monthly values of 52 macroeconomic indicators. Prediction performance of these values from their earlier values was measured with different lags (1 month, 2 months, etc.). The data were processed as described by Carriero et al. (2011). Data for each month were used as a test set (395 test sets) while using data from the previous 10 years for training. Where cross-validation was needed for learning the values of model parameters, data from the last 2 years before the month-to-be-predicted were used for validation and data from the previous 8 years for training.

5.2 Methods included in comparison

We compared the latent-noise BRRR with a state-of-the-art sparse multiple-output regression method Graph-guided Fused Lasso ('GFlasso') (Kim and Xing, 2009), BRRR/factor regression model (Marttinen et al., 2014) with and without the *a priori* independent noise model ('independent-noise BRRR', 'BRRR without noise model'), standard Bayesian linear model ('blm') (Gelman et al., 2004), standard ridge regression ('ridge regression') (Hoerl and Kennard, 1970), elastic-net-penalized multi-task learning ('L2/L1 MTL'), kernel regression with linear and Gaussian kernels combined with a process for removing confounding factors (Stegle et al., 2012) ('KRR with linear kernel + PEER', 'KRR with Gaussian kernel + PEER') and a baseline method of predicting with target data mean. GFlasso constitutes a suitable comparison as it encourages sharing of information between correlated responses, as our model, but does that within the Lasso-type penalized regression framework without the use of a noise model to explain away the structured noise. L2/L1 MTL is a multitask regression method implemented in the `glmnet` package (Friedman et al., 2010) that allows elastic net regularization. It does not use a noise model to explain away confounders either. The blm method and ridge regression were selected as a simple single-task baselines.

In one of the experiments, on an association study, latent-noise BRRR is compared with independent-noise BRRR and canonical correlation analysis ('cca'), considered the state-of-the-art methods for the detection of multivariate associations (Marttinen et al., 2013, 2014). Additionally, the simple univariate linear model ('lm') is included as it represents the common baseline in association analysis.

We compare latent-noise BRRR also with two other new models for structured noise modeling. In the simulations, we study the performance of correlated Bayesian reduced

rank regression ('correlated BRRR'), which is presented in more detail in the Supplementary material. In brief, in the correlated BRRR, the correlation structure of the target variables learnt by an *a priori* independent noise model is used as a prior for the regression weight parameters. With the NFBC1966 data and the macroeconomic time series data sets, we also study the performance of the method presented in Equation (10) in Section 3.3 that explicitly models both latent and independent structured noise, abbreviated as 'latent+independent-noise BRRR'.

Parameters for the different methods were specified as follows:

GFlasso: The regularization parameters of the `gw2` model were selected from the default grid using cross-validation. The method has been developed for genomic data indicating the default values should be appropriate. However, for NFBC1966 data, we were unable to run the method with the smallest values of the regularization parameters $\{110, 60, 10\}$ due to lengthy runtime with these values. With this computational compromise of leaving out these three values, the average training time for the largest training data sets was $\sim$ 650 h. With NFBC1966 data, the pre-specified correlation network required by the GFlasso was constructed to match the VLDL, IDL, LDL, and HDL metabolite clusters from Inouye et al. (2012). Within these clusters, the correlation network was fixed to the empirical correlations, and to 0 otherwise. With DILGOM data, we used the empirical correlation network, with correlations below 0.8 fixed to 0 to reduce the number of edges in the network for computational speedup.

independent-noise BRRR, BRRR without noise model: Hyperparameters a_1 and a_2 of all the BRRR models were fixed to 10 and 4, respectively. In total 1,000 MCMC samples were generated and 500 were discarded as burn-in. In preliminary tests similar results were obtained with 50,000 samples. The remaining samples, thinned by a factor of 10, were used for prediction. The maximum rank of the infinite-rank BRRR model was learned using cross-validation from the set of values $\{5, 10, 15\}$ for the NFBC1966 data set, $\{2, 4, 8\}$ for the metabolomics prediction task on the DILGOM data set and $\{2, 5, 10, 20\}$ for the gene expression prediction task on the DILGOM data set. These grids were selected based on initial experiments. For the fMRI response prediction, the possible values for the maximal rank were limited to $\{2, 4\}$ in order to save computational time. For the econometrics data set, maximum ranks of $\{5, 10, 20\}$ were used. In the association detection task, the rank of independent noise BRRR was fixed to 1 as this was already sufficient for the task.

latent-noise BRRR: With the NFBC1966 data, the latent signal-to-noise ratio β was selected using cross-validation from a range of values from 100 to $\frac{1}{100}$, $\beta = \{100, 10, 2, 1, \frac{1}{7.5}, \frac{1}{15}, \frac{1}{30}, \frac{1}{60}, \frac{1}{100}\}$, in order to thoroughly evaluate the sensitivity of the model to this parameter. For the other data sets and tasks, the sets of values were as follows: DILGOM metabolomics prediction: $\beta = \{10, 2, 1, \frac{1}{7.5}, \frac{1}{15}, \frac{1}{30}, \frac{1}{60}, \frac{1}{100}\}$, DILGOM gene expression prediction $\beta = \{10, 1, \frac{1}{5}, \frac{1}{10}, \frac{1}{30}, \frac{1}{50}, \frac{1}{100}, \frac{1}{300}\}$ and for macroeconomic time series prediction $\beta = \{10, 2, 1, \frac{1}{7.5}, \frac{1}{15}, \frac{1}{30}, \frac{1}{60}, \frac{1}{100}\}$. For fMRI response prediction, the set of values was limited to $\beta = \{10, 1, \frac{1}{10}\}$ to save computation time. Other parameters, including the number of iterations, were set as for the independent-noise BRRR. The performance of the model was evaluated both by sampling the

maximum rank and by learning it with cross-validation from the same range of values as with independent-noise BRRR. Shrinkage hyperparameters were set to non-informative values, a_1 =10 and $a_2 = 4$, similarly to the corresponding parameters a_3 and a_4 of independent-noise BRRR.

blm: The variance hyperparameter of BLM was integrated over using MCMC. The variance hyperparameter was assigned a Gamma prior with both shape and rate parameters set to 1. In total 1,000 posterior samples were generated and 500 were discarded as burn-in.

ridge regression: Ridge regression was used as implemented in the `glmnet` package with default parameters. The default convergence threshold parameters of `glmnet` were used and no warnings/numerical problems occurred.

L1/L2 MTL: The effects of different types of regularization penalties are an active research topic and we ran a continuum of mixtures of L1 and L2 penalties ranging from group lasso to ridge regression. The mixture parameter α controlling the balance between L1 and L2 regularization was evaluated on the grid [0, 0.1, ..., 0.9, 1.0] and selected using a 10-fold cross validation. The default convergence threshold parameters of `glmnet` were used and neither warnings nor numerical problems occurred.

KRR with linear kernel + PEER: First, the PEER software (Stegle et al., 2012) was used to remove the effects of confounders using 15 components. Then kernel ridge regression with a normalized linear kernel (Bishop et al., 2006) was applied using the residuals from PEER as the target variables. Kernel ridge regression was regularized according to the standard approach of adding parameter λ to the diagonal elements of the kernel. The value of λ was selected using cross-validation from a set of 10 values ranging from 0.1 to 100, $[10^{-1}, 10^{-0.66}, \ldots, 10^{1.67}, 10^{2}]$. To share information between the different target variables, the approach of using the same kernel for all target variables was adopted.

KRR with Gaussian kernel + PEER: Kernel ridge regression using a Gaussian kernel was used. Regularization and the use of PEER were otherwise similar to KRR with linear kernel + PEER. The radius parameter of the Gaussian kernel was selected using cross-validation from a set of 30 values ranging from 0.001 to 1000, $[10^{-3}, 10^{-2.79}, \ldots, 10^{2.79}, 10^{3}]$

cca: This is the conventional classical canonical correlation analysis that attempts to identify linear combinations of the columns of the input and output matrices that are maximally correlated with each other.

correlated BRRR: Rank and hyperparameters a_1, a_2, a_3 and a_4 were set as with the independent-noise BRRR. This model is presented in detail in Supplementary Section 1.

latent+independent-noise BRRR: With the NFBC1966 data, the hyperparameters a_1, a_2, a_3 and a_4 were set as with the independent-noise BRRR. The latent signal-to-noise ratio β was selected using cross-validation from a range of values from 100 to $\frac{1}{100}$, $\beta = \{100, 10, 2, 1, \frac{1}{7.5}, \frac{1}{30}, \frac{1}{60}, \frac{1}{100}\}$ and the maximum rank was fixed to 10. For the econometrics data set, maximum ranks of $\{5, 10, 20\}$ were used and the signal-to-noise ratio β was selected using cross-validation from the values $\beta = \{10, 2, 1, \frac{1}{5}, \frac{1}{10}, \frac{1}{30}$. For both data sets, the variance parameter of the *a priori* independent noise H was selected from values $\{10^{-6}, 1\}$, value 10^{-6} corresponding to the extreme case of latent-noise BRRR.

5.3 Simulation experiment: impact of the noise model assumptions

In this Section, we study the implications of different noise model assumptions. Performances of models with different noise model assumptions are measured on simulated data sets generated from a continuum of models between the two extremes of assuming either latent noise, or *a priori* independent regression and noise models. The synthetic data are generated according to

$$Y = (X\Psi + \alpha\Omega)\,\Gamma + (1-\alpha)H\Lambda + E, \tag{18}$$

where $\mathrm{vec}(E) \sim \mathcal{N}(0, I_{NK})$ and the parameter $\alpha \in [0, 1]$ defines the proportion of variance attributed to the latent noise versus independent noise. We study a continuum of problems with the values of parameter $\alpha = 0, 0.1, \ldots, 1$. The parameters Γ and Λ are orthogonalized using Gramm-Schmidt orthogonalization. The parameters are scaled so that covariates X explain 3 % of the variance of Y through $X\Psi\Gamma$, the diagonal Gaussian noise $\mathcal{N}(0, I_{NK})$ explains 20 % of the total variance of Y and the structured noise $\alpha\Omega\Gamma + (1-\alpha)H\Lambda$ explains the remaining 77 % of the total variance of Y. The simulation was repeated 100 times and training data sets of 500 and 2000 samples were generated for each replicate. To compare the methods, performance in mean squared error (MSE) of the models learned with each method was compared to that of the true model on a test set of 15 000 samples. The number of covariates was fixed to 30 and the number of dependent variables to 60. Rank of the regression coefficient matrix and structured noise was set to 3 when simulating the data sets.

For independent-noise BRRR, the rank of the regression coefficient parameters Ψ and Γ was fixed to the true value while the rank of the noise model was learnt from the data. For latent-noise BRRR, the performance of the model was evaluated both by fixing the rank of the regression coefficient matrix to its true value and by learning it from the data. The variance of Ω was selected using 10-fold cross-validation. The grid for latent signal-to-noise ratios β was $\beta = \frac{1}{5}$ to $\frac{1}{15}$, $\beta = \{\frac{1}{5}, \frac{1}{7.5}, \frac{1}{10}, \frac{1}{12.5}, \frac{1}{15}\}$. More specifically, $\mathrm{Var}(\mathrm{vec}(\Omega)) = \sigma_\Omega^2 I_{NK}$ where $\sigma_\Omega^2 = \frac{1}{\beta} \times \mathrm{Trace}\,(\mathrm{Var}\,(X))$. The grid was chosen according to the interpretation given in Section 3.2; it corresponds to assuming that the latent noise explains 5 to 15 times the variance explained by the covariates.

Figures 4 (a) and (b) present the results of a simulation study with training sets of 500 and 2000 samples, respectively. When the structured noise is generated according to the conventional assumption of independent signal and noise, the model making the independence assumption (i.e., the independent-noise BRRR) performs equally well to the true

model with both 500 and 2000 samples. However, when the assumption is violated and the proportion of latent noise increases, the performance of the independent-noise BRRR breaks down, whereas the latent-noise BRRR performs consistently well. The method that does not explain away the structured noise at all (BRRR without noise model) is always inferior to the null model with the training set of 500 samples. When the number of training samples is increased to 2000 and the noise is generated according to the latent-noise assumption, the model, however, outperforms even the independent-noise BRRR.Thus, having no noise model is in this case better than having the noise model based on the incorrect independence assumption, which emphasizes the importance of the assumptions on which the noise model is based. Interestingly, with n=2000 the BRRR without noise model is among the best performing methods whereas with n=500 it is clearly the worst, highlighting the fact that the smaller n gets, the more important the right assumptions become.

The latent-noise end of the continuum appears to be more difficult for the methods that do not account for the structured noise (blm, BRRR without noise model). This weak but consistent trend can be seen in Figure 4(b) where the difference between the oracle and these methods increases with the percentage of latent noise. This behaviour is, however, rather intuitive in terms of Equation (18); by rewriting

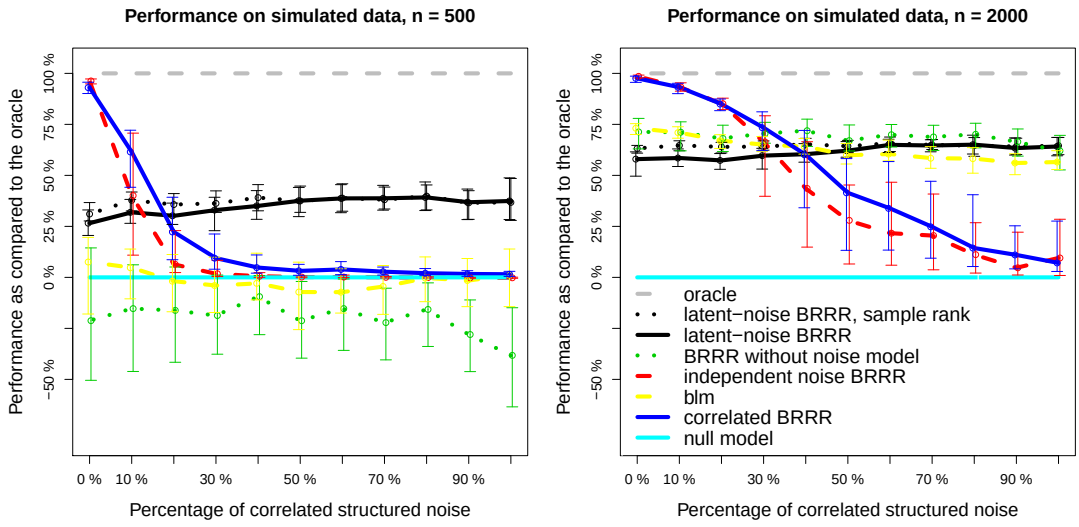
$$\begin{aligned} Y &= (X\Psi + \alpha\Omega)\,\Gamma + (1-\alpha)H\Lambda + E, \\ &= X\Psi\Gamma + \alpha\Omega\Gamma + (1-\alpha)H\Lambda + E \end{aligned}$$

it is obvious that as $\alpha \to 1$, the structured noise (coming from Ω and H) will with certainty be projected on the particular target variables that are affected by the covariates X. In other words, latent noise blurs exactly the relationships of interest, being very disruptive.

Figure 4 shows results also for an alternative novel model that shares information between the noise and regression models (correlated BRRR, see Supplementary material for a detailed description). The model includes a separate noise model for the structured noise, as in (1), but achieves the information sharing by assuming a joint prior for the noise and regression models. In detail, conditional on the noise model, the current residual correlation matrix between the response variables is used as a prior for the rows of Γ. This way the correlations between target variables are propagated into the corresponding regression weights; however, the strongest noise components are not automatically coupled with the strongest signal components. Notably, the performance of the correlated BRRR model is very similar to the regular BRRR model that does not have any dependence between the noise and signal components.

5.4 NFBC1966: metabolomics prediction

In this Section, the models accounting for latent noise are evaluated in terms of predictive performance on the NFBC1966 data with different training set sizes. Figure 5 presents the test data MSE for the different methods. With the larger training set sizes, latent-noise BRRR outperforms the other methods. With the smallest training data size, ridge regression and latent-noise BRRR perform equally outperforming all other methods. However, ridge regression is unable to improve its performance as the number of training data points increases, and with the larger training sets it is outperformed by the more complex methods.



(a) (b)

Figure 4: Performance of different methods, compared to the true model, as a function of the proportion of latent noise with a training set of (a) 500 and (b) 2000 samples. The x-axis indicates the proportion of noise generated according to the latent noise assumptions (100% corresponds to $\alpha = 1$). Bars denote $\pm$ 1 standard deviation, computed independently for each x-coordinate. The performance of 100% means the amount of variance explained by the model is equal to the amount explained by the true model. The performance of 0% means that the method does not explain any variance of the target variables, whereas negative values indicate the variance actually increases after taking the predictions into account.

Method blm performs worse than the baseline (null model, prediction with training set mean), even with the largest training data set containing 3761 individuals, and BRRR without noise model requires the largest training set size in order to outperform the baseline. A paired t-test for the performance difference between latent-noise BRRR and independent-noise BRRR yields a p-value of 0.03 suggesting a statistically significant difference.

5.5 Differences between latent-noise BRRR and independent-noise BRRR on NFBC1966 metabolomics prediction

The two differences between our new approach, latent-noise BRRR, and independent-noise BRRR are (1) model structure (latent-noise BRRR shares parameters between the regression and noise models) and (2) using the latent signal-to-noise ratio parameter β to regularize the model. In order to identify how these developments lead to the observed performance differences on the NFBC1966 data, we performed a sensitivity analysis for the two methods with respect to the assumed amount of variance attributed to the noise model.

Figure 6 presents the results of this sensitivity analysis. For latent-noise BRRR, the assumed variance of the noise model controlled by the *a priori* signal-to-noise ratio β affects

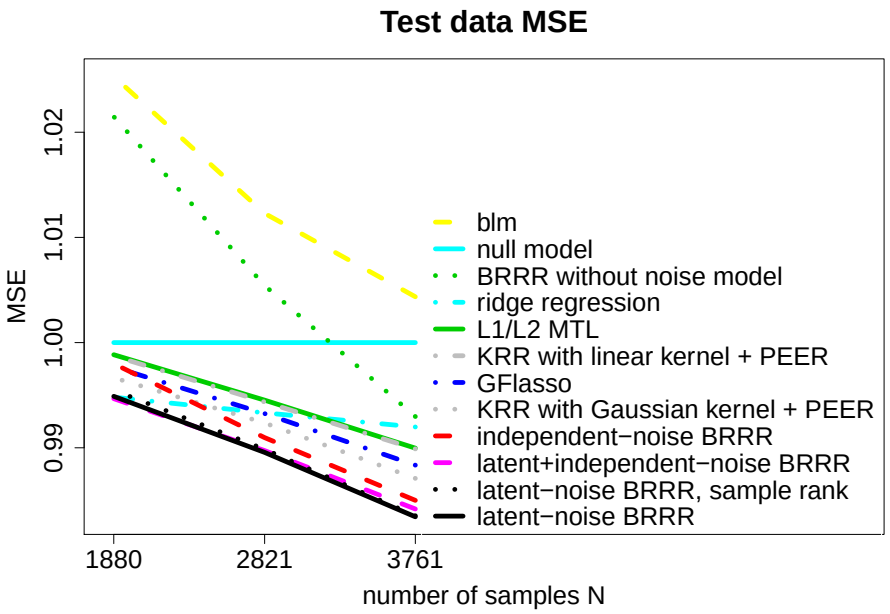


Figure 5: Test data MSE for different amounts of training data on the NFBC1966 metabolomics data. The MSEs have been scaled to give the null model a MSE of 1.

performance in a consistent way, whereas for independent-noise BRRR the impact appears random. If the performance difference stemmed mainly from controlling the variance of the noise model, controlling that parameter for both models should lead to similar results. On the other hand, if the difference in the model structure alone sufficed to explain the performance difference, the difference should not be sensitive to the variance of the noise model. Hence, we conclude that, on this data set, both the new model structure and regularization by using the latent signal-to-noise ratio are required for improved performance.

We also studied the variability of the estimated latent SNR on different folds. The optimal l-SNR was estimated very consistently, the results are presented in Supplementary Figure 3.

5.6 Evaluation of the chosen inference procedures for rank and noise parameters

Inference for the proposed model could naturally be done in several alternative ways. In this Section we justify the proposed inference procedure.

In the simulations (Section 5.3), sampling the maximum rank of the infinite prior worked well, measured in terms of predictive performance. Figure 4 shows that sampling the maximum rank actually improves performance, as compared to fixing it to the value used in the generative process, when the latent noise assumption is wrong (left end), both when $N = 500$ and when $N = 2000$. When the latent noise assumption holds (right end), the two

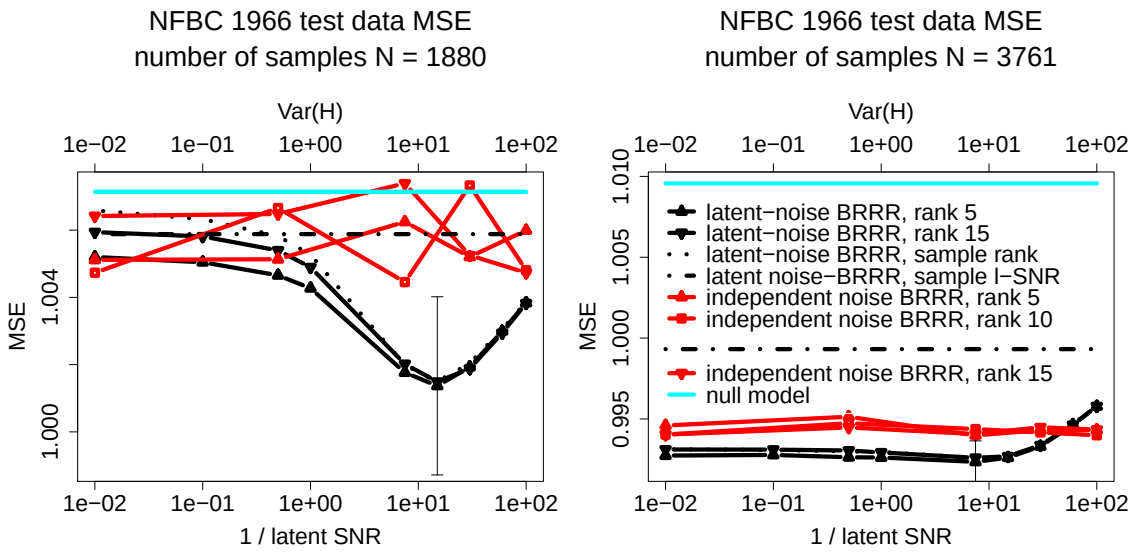


Figure 6: Sensitivity of latent-noise BRRR and independent-noise BRRR to the variance of the structured noise with different maximum ranks. The results are on NFBC1966 test data MSE (N =1880 and N =3761) as a function of the noise model variance. Lower axis: *a priori* latent signal-to-noise ratio of latent-noise BRRR and the upper axis: variance of the model parameter H of independent-noise BRRR. The bar denotes the standard deviation of the test set performance difference observed between the two models in cross-validation. The figures also present the unscaled performance of the null model and the performance of the latent-noise BRRR when using sampling to infer the latent signal-to-noise ratio (latent-noise BRRR, sample l-SNR) and when using sampling to infer the maximum rank (latent-noise BRRR, sample rank). When $N = 3761$, sampling the rank results in similar performance as obtained with the fixed values and thus the curves overlap.

inference procedures perform equally well. With the NFBC1966 data set (Figure 5), learning the maximum rank of the infinite prior by sampling (latent-noise BRRR, sample rank) or by cross-validation (latent-noise BRRR) results in very similar test set performances, similarly to in the simulation experiment in Section 5.3. Hence, we conclude that for learning the maximum rank, both sampling and cross-validation are appropriate techniques. We also ran the independent-noise BRRR so that the rank was sampled instead of selecting it using cross-validation on this data. However, the results were poor, with the test data MSE equal to 1.019 with N =1880 and 1.005 with N =3761. The lines were omitted for clarity. We hypothesize that the problems with the instability of the model (see Section 5.11) were accentuated when the rank was sampled.

The key parameter of our model, the latent signal-to-noise ratio, was estimated using cross-validation. In the simulations, the cross-validation based scheme allowed estimation of the latent signal-to-noise ratio to a reasonable accuracy. The estimated values are included in Figure 2 in the Supplementary material. While the latent signal-to-noise ratio of the

generative process was $\approx 1/25$, the estimated posterior latent signal-to-noise ratios ranged from $\frac{1}{14}$ to $\frac{1}{19}$ in the parts of the domain where the percentage of correlated structured noise was 100-80%. When the percentage of correlated structured noise was 0-10 %, the model correctly learnt lower variance for the latent noise and a corresponding stronger latent signal-to-noise ratio β.

We also studied the performance of latent-noise BRRR while sampling the variance of the noise model. A non-informative prior was assigned for the variance of Ω, $\Omega \sim \mathcal{N}(0, \sigma_\Omega^2)$ and $\sigma_\Omega^{-2} \sim \text{Gamma}(\text{shape} = 0.001, \text{rate} = 0.001)$. The performance of this model is presented in Figure 6. The performance of latent-noise BRRR when sampling the variance of Ω is consistently worse than when using cross-validation to select the value of the latent signal-to-noise ratio. Hence, we conclude that, as opposed to other parameters, cross-validation is needed to learn the latent-signal to ratio to reach the improved performance.

5.7 NFBC1966: multivariate association detection

Detection of associations between multiple SNPs and metabolites is a topic that has received attention recently (see, e.g., Kim et al., 2009; Inouye et al., 2012; Marttinen et al., 2014). Here we demonstrate the potential of the new method in this task using two illustrative example genes for which ground truth is available. Associations between SNPs within two genes, *LIPC* and *XRCC4*, and the metabolites in the NFBC1966 data are investigated in the experiment. Note that the covariates (SNPs) used in this experiment are different from the ones used in the prediction experiment: here SNPs in individual genes are used, whereas in the prediction experiment all known lipid-associated SNPs were used. *LIPC* was selected as a reference, because it is one of the most strongly lipid-associated genes. On the contrary, *XRCC4* was discovered only recently using three cohorts of individuals (Marttinen et al., 2014), and it was selected to serve as an example of a complex association detectable only by associating multiple SNPs with multiple metabolites, and not visible using simpler methods.

We use the proportion of total variance explained (PTVE) as the test score (Marttinen et al., 2014), and sample 100 permutations to measure the power to detect the associations. Furthermore, we use downsampling to evaluate the impact of the amount of training data. For comparison, we select the BRRR, the exhaustive pairwise (univariate) linear regression ('lm'), and canonical correlation analysis (CCA) (Ferreira and Purcell, 2009), these being the methods that have been proposed for the task and having a sensible runtime in putative genome-wide applications. For lm, the minimum p-value of the regression coefficient over all SNP-metabolite pairs, and for the CCA, the minimum p-value over all SNPs (each SNP associated with all metabolites jointly) are used as the test scores. The association involving the *XRCC4* gene was originally detected using the BRRR model; however, unlike here, informative priors were used for the regression coefficients.

Table 1 presents the ranking of the original data among the permuted data with different sample sizes and methods. Ten MCMC chains were computed for both models to account for sampling variability on this difficult and relatively strongly collinear data. The association score was obtained by averaging over the scores for different chains. As expected, all methods were able to detect the association involving *LIPC* with both training set sizes. However, latent-noise BRRR had the highest power to detect the *XRCC4* gene.

Table 1: Power of different methods to detect the association between metabolomics profiles and *XRCC4* or *LIPC* genes with N =4702 and N =2351 samples. Power is measured as the proportion of association test scores in permuted data sets smaller than the test score in the original data set. Value 1 indicates that the association score of the unpermuted data was higher than the score in any permutation.

	XRCC4		*LIPC*	
	N =4702	N =2351	N =4702	N =2351
latent-noise BRRR	0.98	0.94	1	1.00
independent-noise BRRR	0.41	0.32	1	0.99
lm	0.62	0.74	1	1.00
cca	0.20	0.24	1	1.00

5.8 Results: other real-world data sets

To thoroughly study the empirical value of the new method, we compared it to alternative methods on macroeconomic time series prediction, metabolomics and gene expression prediction experiments on the DILGOM data set and the fMRI response prediction. In these domains, explaining away structured noise is of crucial importance.

With the DILGOM data, the prediction of the weak effects was challenging for all methods. Indeed, we noticed that the null model using the average training data value for prediction was better than any other method in terms of MSE over all target variables with the single exception of L1/L2 MTL, which set all regression coefficients to zero thus reducing to the null model. However, a detailed investigation of the results revealed that while many of the target variables could not be predicted at all (as indicated by the worse than null model MSE) some of the target variables could still be predicted better than the null model, and by focusing the analysis on the MSE computed over the predictable target variables (*i.e.,* those that could be predicted better than the null by at least one method), comparisons regarding the model performances could still be made. For consistency, both metrics were computed also with the fMRI and econometrics data sets. To save computation time, we chose to evaluate only the cross-validation based variant of our model for the fMRI data, as this approach had already been identified as the most promising implementation of our method.

Table 2 and supplementary Table 1 present the results of the macroeconomic time series prediction experiment, metabolomics and gene expression prediction experiments on the DILGOM data set and the fMRI response prediction experiments. The results have been normalized so that the score for the null model (prediction using the mean) is 1. Table 2 presents the results for the predictable target variables and supplementary Table 1 presents the results obtained by averaging test data MSE over all target variables.

Latent-noise BRRR outperforms independent-noise BRRR consistently on the gene expression (on 8/10 folds), metabolomics (10/10 folds) and fMRI response prediction (2/2 folds) tasks on both scores. In the fMRI response prediction, the latent-noise BRRR and L1/L2 MTL are the only methods that outperform the null model. With the DILGOM data none of the methods outperformed the null model when averaged over all target variables and when concentrating on the predictable target variables, only the latent-noise BRRR and

KRR with Gaussian kernel were able to outperform the null model. With gene expression prediction, latent noise BRRR (sample rank), GFlasso and the kernel methods outperform the null model, latent-noise BRRR being the best. The econometrics data is the only case in which the independent-noise BRRR is more accurate than the latent-noise BRRR on both metrics, and the latent-noise BRRR is the third best method. In this data set, however, the effects appear rather strong as different methods explain up to 10-32% of the variance of the target variables and the best method is, in fact, ridge regression.

On the small DILGOM data sets, L1/L2 MTL sets all regression weights to zero as hypothesized in Section 3.2. This demonstrates the need to develop new alternatives to L1/L2 regularization: when modeling weak effects on small data sets, using L1/L2 penalties can prevent analysis altogether. Ridge regression appears to suffer from the same problem on the NFBC data set: although shrinking weights towards zero efficiently avoids overfitting, the model is only able to learn the strongest effects. Thus ridge regression outperforms most methods on the smallest training set (where the complex methods easily overfit), but heavily loses as the training set increases and the more complex methods become able to also benefit from the weak effects. Regularization by making the noise model stronger as in latent-noise BRRR avoids this problem.

The standard method blm performs surprisingly poorly especially as compared to ridge regression. The implementation was checked carefully. Predictive performance with the standard least squares linear model was also evaluated for some of the data sets (results not shown) and we found that it performed even worse than the blm. We hypothesize that the collinearity present in all of the data sets analyzed harmed the performance of blm and lm more than that of ridge regression.

5.9 Results: simultaneous modeling of both latent and independent structured noise

As both latent and independent structured noise can be present simultaneously, we evaluated the possible gains from taking both noise types simultaneously into account. A model that incorporates both latent and independent structured noise, here called latent+independent-noise BRRR, was evaluated for the metabolomics prediction task on the NFBC1966 data and on the macroeconomic time series prediction task, the strong domains of the methods of interest.

Results of this experiment are presented in Table 3. In metabolomics prediction, accounting for both noise types improved results slightly on the smallest training data size as compared to the best performing method latent-noise BRRR. On the larger training data sets, the more flexible latent+independent-noise BRRR model performed worse than the latent-noise BRRR that only accounts for latent noise. On the macroeconomic time series prediction task, accounting for both noise types improved performance as compared to only accounting for the dominant noise type (independent structured noise) on the smaller training data set. For summary,even though slight performance improvements were seen with the smallest training set sizes, the results indicate that as the size of the training data set increases, the advantages disappear. We hypothesize that the potential under-identifiability issues discussed in Section 3.3 hinder model performance more than the increased flexibility improves it.

	econometrics	DILGOM: gene expression	DILGOM: metabolomics	fMRI
latent-noise BRRR	**0.73320±0.22564**	**0.99990±0.00057**	1.00046±0.00130	**0.99798±0.00282**
latent-noise BRRR, sample rank	**0.73453±0.21219**	1.00039±0.00107	**0.99995±0.00100**	
independent-noise BRRR	**0.71072±0.20549**	1.00051±0.00038	1.04163±0.03781	1.00215±0.00183
L1/L2 MTL	**0.75035±0.15651**	1.00000±0.00000	1.00000±0.00000	**0.99786±0.00090**
GFlasso		1.00010±0.00106	**0.99996±0.00221**	
KRR with linear kernel + PEER	**0.88138±0.11021**	1.00093±0.00057	**0.99995±0.00006**	1.00236±0.00112
KRR with Gaussian kernel + PEER	**0.90497±0.09707**	**0.99985±0.00016**	**0.99998±0.00004**	1.00649±0.00179
BRRR without noise model	**0.81818±0.34747**	1.00568±0.00274	1.30795±0.08802	1.06722±0.05586
ridge regression	**0.689771 ±0.202603**	1.001798±0.001090	1.000445±0.003089	1.003388±0.007420
blm	1.59040±1.23041	1.04245±0.00914	1.52573±0.08859	2.08650±0.25396
null model	1.00000±0.00000	1.00000±0.00000	1.00000±0.00000	1.00000±0.00000

Table 2: Test data MSE computed on the predictable target variables on the econometrics, DILGOM and fMRI data sets. Bold font indicates better than baseline accuracy achieved by predicting with the training data mean.

5.10 Improvement in computational efficiency resulting from the reparameterization of model

To confirm the computational speed-up resulting from the reparameterization presented in Section 4, we performed an experiment where the algorithm implementing the naïve

	NFBC $N = 3761$	NFBC $N = 1880$	econometrics $N = 120$	econometrics $N = 60$
latent-noise BRRR	**0.9833±0.0077**	0.9949±0.0037	0.7536±0.2143	0.8374±0.1816
latent+independent-noise BRRR	0.9840±0.0072	**0.9947±0.0019**	0.7445±0.1918	**0.7889±0.1561**
independent-noise BRRR	0.9849±0.0078	0.9980±0.0059	**0.7339±0.1977**	0.8097±0.2085

Table 3: Performance of the most flexible modeling assumptions. Test data MSE on the NFBC and econometrics data sets. On the larger training data sets, latent-noise BRRR and independent noise BRRR outperform the model that accounts for both noise types, latent+independent-noise BRRR. On the smaller training data sets, however, this model outperforms the models that only account for one noise type.

Gibbs sampling updates for the Bayesian reduced-rank regression (Geweke, 1996; Karlsson, 2012) was compared with the new algorithm that uses the reparameterization. Similar improvements were achieved with all other BRRR models as well.

Ten simulated data replicates were generated from the prior. The number of samples in the training set was fixed to 5000 and the number of target variables was set to 12. Rank of the regression coefficient matrix was 2. Runtime was measured as a function of the number of covariates, which was varied from 100 to 300; 1000 posterior samples were generated. The new algorithm that reparameterizes the model clearly outperformed the naïve Gibbs sampler (Figure 7). As a sanity check, the regression coefficient matrices estimated by the algorithms were compared , and found to be similar.

5.11 Efficiency of the algorithm

To investigate the efficiency of the proposed algorithm and to compare it with the alternative methods, we recorded the wall-clock run times with the NFBC1966 data set, shown in Figure 8. In addition, we studied the conventional convergence diagnostics. To assess convergence and mixing, we re-computed four MCMC chains of 2000 posterior samples each, for each of the BRRR methods. Averaged effective sample sizes (ESS) and potential scale reduction factors (PSRF) were computed for 200 randomly selected parameters of the regression coefficient matrix (Gelman et al., 2004). These results are presented in Table 4.

All BRRR methods, except for independent-noise BRRR, converge (PSRF < 1.1) and mix acceptably efficiently ($\frac{N_{\text{effective}}}{N_{\text{samples}}} \approx \frac{40}{1000}$). Independent noise BRRR, however, showed poor mixing and convergence. In initial experiments we observed that the PSRF for the independent-noise BRRR did not necessarily ever reach values indicating convergence even when sampled for 15,000 iterations. Thus, we decided to simply use the same number of MCMC iterations for each method in our experiments. The reason for the bad behaviour was the multimodality of the posterior distribution, caused by the too flexible model structure of the independent noise model, and the resulting convergence of the different chains into different modes.

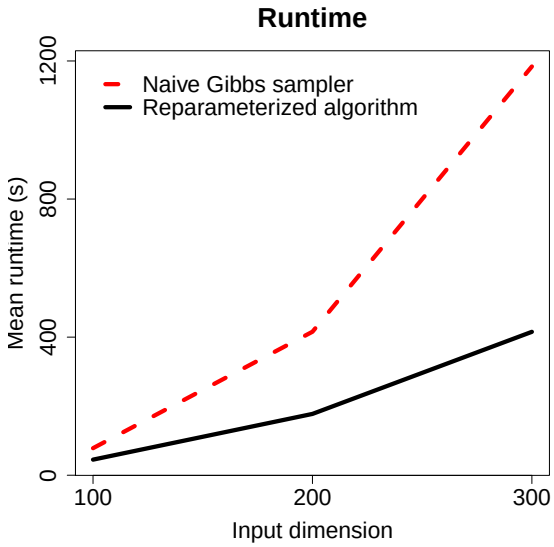


Figure 7: Runtime of the algorithm implementing the naïve Gibbs sampler with computational complexity and the new algorithm that reparameterizes the model. The naïve algorithm has a computational complexity of $O(P^3 S_1^3)$ and the new algorithm $O(P^3 + S_1^3)$. Random variation over the repetitions was minimal and the error bars were omitted for clarity.

	independent-noise BRRR	BRRR without noise model	latent-noise BRRR	latent-noise BRRR, sample rank
1000 samples	4.46 ± 0.32	1.03 ± 0.03	1.06 ± 0.05	1.01 ± 0.004
2000 samples	3.42 ± 0.18	1.02 ± 0.02	1.05 ± 0.05	1.01 ± 0.003

Table 4: Averaged PSRF.

To further demonstrate the difference between the latent-noise and independent-noise BRRR methods, we visualized the MCMC trace of the association metric used in Section 5.7. The instability of independent-noise BRRR is strikingly visible in Figure 9. The chains converge to different modes and mix very slowly. On the other hand, the latent-noise BRRR appears to mix adequately and always converges to the same mode, except for one of the ten chains with the XRCC4 gene, which converges to a mode with a lower value of the explained variance.

5.12 Results: summary of the results with the real data sets

To provide an overview of the performances of the different methods on the various data sets and tasks, the methods' performances were ranked for each task/data set. For the prediction tasks, methods were ranked according to the MSE on the test set. When none of the methods outperformed the null model, the scores on the predictable target variables

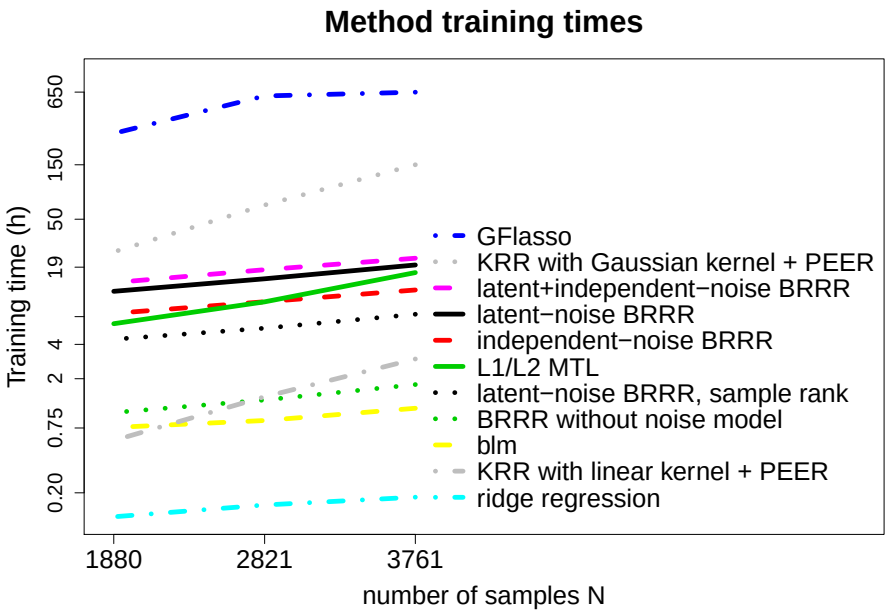


Figure 8: Computation times of the methods for different training set sizes N on the NFBC1966 metabolomics data.

	independent-noise BRRR	BRRR without noise model	latent-noise BRRR	latent-noise BRRR, sample rank
1000 samples	4.32 ± 0.32	43.88 ± 0.93	44.83 ± 0.25	40.74 ± 1.18
2000 samples	5.15 ± 0.66	84.39 ± 1.61	86.40 ± 0.43	77.77 ± 1.25

Table 5: Effective sample sizes for the Bayesian reduced rank regression methods. Independent-noise BRRR mixes substantially worse than the other methods.

were compared instead. In the association detection task, estimated statistical power was used as the ranking criterion. Table 6 presents the overview results.

Averaged over all data sets and tasks, latent-noise BRRR outperforms the comparison methods. In particular, the latent-noise BRRR outperforms the independent-noise BRRR on all setups except for the macroeconomic time series prediction task, where independent-noise BRRR is the best method and the two variants of latent-noise BRRR follow. The difference between the latent-noise BRRR and the independent-noise BRRR is consistent, present on 4/5 test folds on the NFBC1966 metabolite prediction, 8/10 test folds on the DILGOM gene expression prediction, 10/10 test folds on DILGOM metabolite prediction and on 2/2 folds on the fMRI response prediction. On macroeconomic time series prediction, independent-noise BRRR is better on 218/395 test folds. In the association detection task the latent-noise BRRR has higher power with both training set sizes on the challenging XRCC4 gene (0.94 vs. 0.32 with n=2,351; 0.98 vs. 0.41 with n=4,702).

Simultaneously accounting for both latent and independent structured noise improves performance on the smallest training data sets considered in the macroeconomic time series prediction and metabolomics prediction (NFBC1966) as compared to accounting for only one type of noise. On the other hand, with the larger training set sizes, the models with just the dominant noise type present perform better than the model including both noise types simultaneously.

Selecting the rank for latent-noise BRRR by sampling or by cross-validation results in comparable performance. Average performance ranks for cross-validation based and sampling-based inferences are 2.5 and 3.5, respectively. For the NFBC1966 data set and gene expression prediction task on the DILGOM data set, cross-validation yields better performance. On metabolomics prediction on DILGOM and the macroeconomic time series prediction, on the other hand, the sampling-based approach works better. It is also intriguing that similarly to the simulations, the sampling based variant of the model works better with independent structured noise (macroeconomic time series prediction) than the cross-validation based approach.

Latent-noise BRRR outperforms the null model on all test cases except for the metabolomics prediction on the DILGOM data. Even on that data set, however, the variant of the model that samples the maximum rank of the infinite prior outperforms the null model. We hypothesize that the poor performance may have resulted from convergence to some inferior mode of the posterior distribution; this can happen to latent-noise BRRR (as demonstrated in Figure 9) although the sharing of information between the signal and noise models makes it substantially more stable than the independent-noise BRRR.

6. Discussion

In this work, we evaluated the performance of multiple-output regression with different assumptions for the structured noise. While most existing methods assume *a priori* independence of the interesting effects and the uninteresting structured noise, we started from the opposite assumption of strong dependence between the components of the model. This assumption may be deemed appropriate for instance with the molecular biological data sets often analyzed with such methods. Using simulations we demonstrated the harmfulness of the independence assumption when latent noise was present. In real data experiments the model assuming latent noise outperformed state-of-the-art methods in prediction of metabolite measurements from genotype (SNP) data and fMRI response prediction, and showed consistently good performance in the different domains. In an illustrative multivariate association detection task, the latent noise model had increased power to detect associations invisible to other methods. To better address the computational needs, we presented a new algorithm reducing the runtime considerably, and improving the scalability of the BRRR models as the number of variables increases. The prior distributions were parameterized in terms of the new concept of *latent signal-to-noise ratio*, which was a key ingredient for optimal model performance. In addition, the rotational unidentifiability of the model was solved using ordered infinite-dimensional shrinkage priors. We also demonstrated that the two modifications (model structure, regularization through the latent signal-to-noise ratio) made to the existing state-of-the-art noise modeling approach were both needed in order to reach the optimal performance.

	NFBC $N = 3761$	econometrics $N = 120$	DILGOM: gene expression $N = 458$	DILGOM: metabolomics $N = 458$	fMRI $N = 1307$	NFBC: XRCC4 association detection $N = 4702$	Average rank
latent-noise BRRR	**1**	3	2	8	2	**1**	**2.8**
latent-noise BRRR, sample rank	2	4	6	**1**			3.2
independent noise BRRR	3	2	7	9	4	3	4.7
L1/L2 MTL	7	5	4	5	**1**		4.4
GFlasso	4		5	3			4.0
KRR with linear kernel + PEER	6	7	8	2	5		5.6
KRR with Gaussian kernel + PEER	5	8	**1**	4	7		5.0
BRRR without noise model	9	6	10	10	8		8.6
blm	11	10	11	11	9		10.4
null model	10	9	3	6	3		6.2
ridge regression	8	**1**	9	7	6		6.2
cca						4	
lm						2	

Table 6: Summary: ranking of methods according to performance in each studied data set and task.

In real data both latent and independent structured noise can be present. We studied a model incorporating both types simultaneously, and, based on these results, we concluded that the possible gains in predictive power as compared to modeling only the dominant type of noise were not worthwhile. In fact, results were also found to degrade when both noise types were included, which we hypothesize to be the result of poor identifiability of the corresponding model

The new model implementing the concept of latent noise was studied using high-dimensional data containing weak signal (weak effects). The new model exploits a ubiquitous character-

istic of such data: while the interesting effects are weak, the noise is strong. Latent-noise BRRR borrows statistical strength from the noise model so as to alleviate learning of the weak effects, by automatically enforcing the regression coefficients on correlated target variables to be correlated. This intuitive characteristic can be seen as a counterpart of the powered correlation priors (Krishna et al., 2009) in the target variable space: Krishna et al. used the correlation structure of the covariates as a prior for the regression weights to enforce correlated covariates to have correlated weights.

The latent-noise BRRR is an extension of several common model families. By removing the covariates, the model reduces to a standard factor analysis model, which explains the output data with underlying factors. Thus, the latent-noise BRRR can be seen as a reversed analogy of PCA regression (West, 2003), in which components of the input space are used as covariates in prediction; in latent-noise BRRR components derived from the output space are predicted using the covariates (see Bo and Sminchisescu, 2009). Allowing the noise term to affect the latent space directly results in interesting connections to *linear mixed models* (LMMs) and *best linear unbiased prediction* (BLUP) (Robinson, 1991); using the latent noise formulation, the model can explain away bias in the residuals as in BLUP. On the other hand, LMMs have a random term for each sample and target variable. While LMMs are not computationally feasible to generalize for high-dimensional targets due to the NK random effect parameters and the associated inversion of an $NK \times NK$ covariance matrix, the latent-noise BRRR can be seen as a low-rank generalization of LMMs for high-dimensional target variables: the covariates are used for prediction in the latent space and in this space there is a noise term for each sample and dimension. Therefore, the number of random effect parameters stays at NS_1 and inference remains tractable.

In summary, our findings extend the existing literature on modeling structured noise in an important way by showing that structured noise can, and should, be taken advantage of when learning the interesting effects between the covariates and the target variables, and how this can be done. Code in R for the new method is available for download at `//http://research.cs.aalto.fi/pml/software/latentNoise/`.

Acknowledgments

This work was financially supported by the Academy of Finland (the Finnish Centre of Excellence in Computational Inference Research COIN; grant numbers 259272 and 286607 to PM; grant number 257654 to MP; grants numbered 140057, 294238 and 292334 to SK).

This work was also supported by the "Machine Learning for Augmented Science and Knowledge Work" project of Aalto University funded by Tekes the Finnish Funding Agency for Innovation (Dnro 1718/31/2014).

NFBC1966 received financial support from the Academy of Finland (project grants 104781, 120315, 129269, 1114194, 24300796, Center of Excellence in Complex Disease Genetics and SALVE), University Hospital Oulu, Biocenter, University of Oulu, Finland (75617), NHLBI grant 5R01HL087679-02 through the STAMPEED program (1RL1MH083268-01), NIH/NIMH (5R01MH63706:02), ENGAGE project and grant agreement HEALTH-F4-2007-201413, EU FP7 EurHEALTHAgeing -277849 and the Medical Research Council, UK (G0500539, G0600705, G1002319, PrevMetSyn/SALVE).

The development and applications of the quantitative serum NMR metabolomics platform are supported by the Academy of Finland, the Sigrid Juselius Foundation, Strategic Research Funding from the University of Oulu, the British Heart Foundation, the Wellcome Trust and the Medical Research Council, UK.

We acknowledge the computational resources provided by the Aalto Science-IT project.

Disclosure: AJK, PS and MAK are shareholders of Brainshake Ltd., a company offering NMR-based metabolite profiling.

References

A. Bargi, R. Y. Xu, Z. Ghahramani, and M. Piccardi. A non-parametric conditional factor regression model for multi-dimensional input and response. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33, pages 77–85. JMLR W&CP, 2014.

J. Baxter. A Bayesian/information theoretic model of bias learning. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, COLT '96, pages 77–88, New York, NY, USA, 1996. ACM.

A. Bhattacharya and D. Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98 (2):291–306, 2011.

C. M. Bishop et al. *Pattern recognition and machine learning*, volume 1. Springer, 2006.

L. Bo and C. Sminchisescu. Supervised spectral latent variable models. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5, pages 33–40. JMLR W&CP, 2009.

L. Bottolo, E. Petretto, S. Blankenberg, F. Cambien, S. Cook, L. Tiret, and S. Richardson. Bayesian detection of expression quantitative trait loci hot spots. *Genetics*, 189(4):1449–1459, 2011.

L. Breiman and J. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54, 1997.

M. Calus and R. Veerkamp. Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution*, 43(1):26, 2011.

A. Carriero, G. Kapetanios, and M. Marcellino. Forecasting large datasets with Bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, 26(5):735–761, 2011.

R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

O. Davis, G. Band, M. Pirinen, C. Haworth, E. Meaburn, Y. Kovas, N. Harlaar, S. Docherty, K. Hanscombe, M. Trzaskowski, et al. The correlation between reading and mathematics ability at age twelve has a substantial genetic component. *Nature Communications*, 5, 2014.

A. Evgeniou and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems 19*, volume 19, pages 41–48, Cambridge, MA, 2007. The MIT Press.

M. Ferreira and S. Purcell. A multivariate test of association. *Bioinformatics*, 25(1):132–133, 2009.

R. Foygel, M. Horrell, M. Drton, and J. Lafferty. Nonparametric reduced rank regression. In *Advances in Neural Information Processing Systems 25*, pages 1637–1645, 2012.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

N. Fusi, O. Stegle, and N. Lawrence. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Computational Biology*, 8(1):e1002330, 2012.

A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman & Hall/CRC, 2004.

J. Geweke. Bayesian reduced rank regression in econometrics. *Journal of Econometrics*, 75 (1):121–146, 1996.

Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45(11):1274–1283, 2013.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

M. Inouye, J. Kettunen, P. Soininen, K. Silander, S. Ripatti, L. S Kumpula, E. Hämäläinen, P. Jousilahti, A. J. Kangas, S. Männistö, et al. Metabonomic, transcriptomic, and genomic variation of a population cohort. *Molecular Systems Biology*, 6(1), 2010.

M. Inouye, S. Ripatti, J. Kettunen, L. Lyytikäinen, N. Oksala, P. Laurila, A. Kangas, P. Soininen, M. Savolainen, J. Viikari, et al. Novel loci for metabolic networks and multitissue expression studies reveal genes for atherosclerosis. *PLoS Genetics*, 8(8):e1002907, 2012.

H. M. Kang, C. Ye, and E. Eskin. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4): 1909–1925, 2008.

S. Karlsson. Conditional posteriors for the reduced rank regression model. Technical Report Working Papers 2012:11, Orebro University Business School, 2012.

J. Kettunen, T. Tukiainen, A-P. Sarin, A. Ortega-Alonso, E. Tikkanen, L-P. Lyytikäinen, A. J. Kangas, P. Soininen, P. Würtz, K. Silander, et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics*, 44 (3):269–276, 2012.

S. Kim and E. Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genetics*, 5(8):e1000587, 2009.

S. Kim, K. Sohn, and E. Xing. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25(12):i204–i212, 2009.

A. Klami, S. Virtanen, and S. Kaski. Bayesian canonical correlation analysis. *The Journal of Machine Learning Research*, 14(1):965–1003, 2013.

G.M. Koop, Rodney W. Strachan, Herman Van Dijk, Mattias Villani, K. Patterson, and T. Mills. *Bayesian approaches to cointegration*, pages 871–898. 2006. ISBN 1403941556. Working paper version - Department of Economics, University of Leicester: Discussion Papers in Economics number 04/27.

A. Krishna, H. D. Bondell, and S. K. Ghosh. Bayesian variable selection using an adaptive powered correlation prior. *Journal of Statistical Planning and Inference*, 139(8):2665–2674, 2009.

J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

P. Marttinen, J. Gillberg, A. Havulinna, J. Corander, and S. Kaski. Genome-wide association studies with high-dimensional phenotypes. *Statistical Applications in Genetics and Molecular Biology*, 12(4):413–431, 2013.

P. Marttinen, M. Pirinen, A-P. Sarin, J. Gillberg, J. Kettunen, I. Surakka, A. J. Kangas, P. Soininen, P. OReilly, M. Kaakinen, M. Kähönen, T. Lehtimäki, M. Ala-Korpela, O. T. Raitakari, V. Salomaa, M. R. Järvelin, S. Ripatti, and S. Kaski. Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression. *Bioinformatics*, 2014.

P. Rai, A. Kumar, and H. Daume III. Simultaneously leveraging output and task structures for multiple-output regression. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3194–3202. Curran Associates, Inc., 2012.

B. Rakitsch, C. Lippert, K. Borgwardt, and O. Stegle. It is all in the noise: Efficient multi-task gaussian process inference with structured residuals. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1466–1474. Curran Associates, Inc., 2013.

P. Rantakallio. Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatrica Scandinavica*, 193:Suppl–193, 1969.

G. K. Robinson. That blup is a good thing: The estimation of random effects. *Statistical science*, pages 15–32, 1991.

K-A. Sohn and S. Kim. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In N. Lawrence and M. Girolami, editors, *International Conference on Artificial Intelligence and Statistics*, volume 22, pages 1081–1089. JMLR W&CP, 2012.

P. Soininen, A. Kangas, P. Würtz, T. Tukiainen, T. Tynkkynen, R. Laatikainen, M. Järvelin, M. Kähönen, T. Lehtimäki, J. Viikari, et al. High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst*, 134(9):1781–1785, 2009.

O. Stegle, L. Parts, R. Durbin, and J. Winn. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Computational Biology*, 6(5):e1000770, 2010.

O. Stegle, C. Lippert, J. M Mooij, N. D. Lawrence, and K. M. Borgwardt. Efficient inference in matrix-variate gaussian models with iid observation noise. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 630–638. Curran Associates, Inc., 2011.

O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507, 2012.

M. Stephens. A unified framework for association analysis with multiple related phenotypes. *PLoS ONE 8(7): e65245*, 2013.

J. H. Stock and M. W. Watson. Forecasting with many predictors. *Handbook of economic forecasting*, 1:515–554, 2006.

T. Teslovich, K. Musunuru, A. Smith, A. Edmondson, I. Stylianou, M. Koseki, J. Pirruccello, S. Ripatti, D. Chasman, C. Willer, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713, 2010.

S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 457–464, New York, NY, 2011. ACM.

W. Wang, V. Baladandayuthapani, J. Morris, B. Broom, G. Manyam, and K. Do. Integrative Bayesian analysis of high-dimensional multi-platform genomics data. *Bioinformatics*, 29(2):149–159, 2012.

L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE*, 9(11):e112575, 11 2014. doi: 10.1371/journal.pone.0112575. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0112575`.

M. West. Bayesian factor regression models in the large p, small n paradigm. *Bayesian Statistics*, 7:733–742, 2003.

C. Xu, X. Wang, Z. Li, and S. Xu. Mapping QTL for multiple traits using Bayesian statistics. *Genetical Research*, 91(1):23–37, 2009.

N. Yi and S. Banerjee. Hierarchical generalized linear models for multiple quantitative trait locus mapping. *Genetics*, 181(3):1101–1113, 2009.

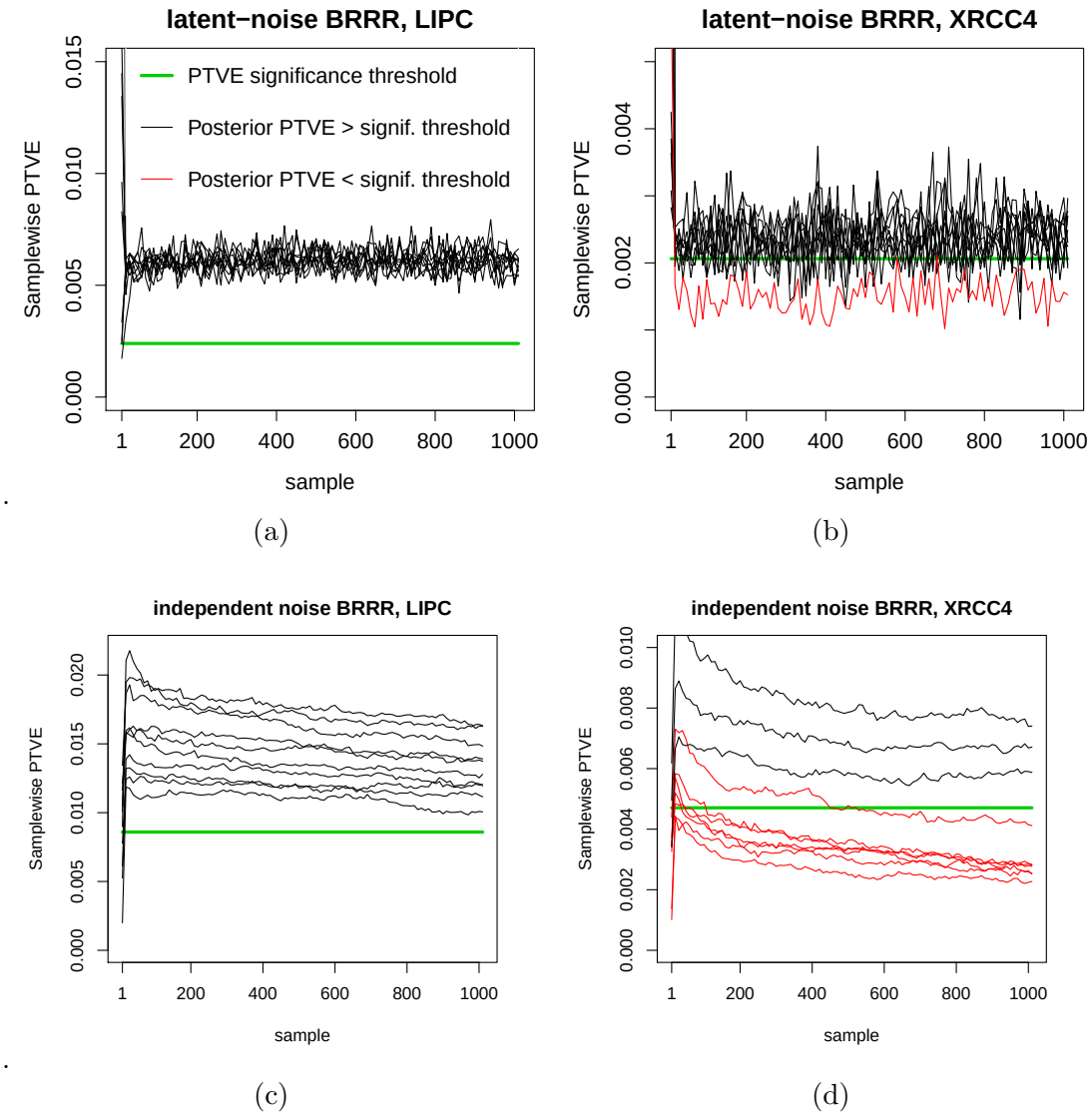


Figure 9: Convergence plots of the association score parameter, that is, the proportion of the total variance explained (PTVE), for latent-noise BRRR and independent-noise BRRR. 10 MCMC chains were computed using data sets with 4702 samples from genes *LIPC* and *XRCC4*. The green line marks the 0.05 significance level of the test score, obtained by permutation sampling. The chains, whose association scores exceed the significance threshold, are drawn in black, whereas the chains that do not exceed it are drawn in red. Latent-noise BRRR converges and mixes appropriately: chains with different initializations converge and traverse the posterior. On the contrary, the independent-noise BRRR behaves rather pathologically: different chains converge to different solutions and explore the posterior poorly.