LETTER **Communicated by Jo-Anne Ting**

Bayesian Sparse Partial Least Squares

Diego Vidaurre
diego.vidaurre@ohba.ox.ac.uk
Oxford Centre for Human Brain Activity, University of Oxford, Oxford OX3 7JX, U.K.

Marcel A. J. van Gerven
m.vangerven@donders.ru.nl
Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behavior, Nijmegen 6525H, Netherlands

Concha Bielza
mcbielza@fi.upm.es
Pedro Larrañaga
pedro.larranaga@fi.upm.es
Computational Intelligence Group, Universidad Politécnia Madrid, Madrid 28660, Spain

Tom Heskes
t.heskes@science.ru.nl
Radboud University Nijmegen, Institute for Computing and Information Science, Intelligent Systems, Nijmegen 6525H, Netherlands

Partial least squares (PLS) is a class of methods that makes use of a set of latent or unobserved variables to model the relation between (typically) two sets of input and output variables, respectively. Several flavors, depending on how the latent variables or components are computed, have been developed over the last years. In this letter, we propose a Bayesian formulation of PLS along with some extensions. In a nutshell, we provide sparsity at the input space level and an automatic estimation of the optimal number of latent components. We follow the variational approach to infer the parameter distributions. We have successfully tested the proposed methods on a synthetic data benchmark and on electrocorticogram data associated with several motor outputs in monkeys.

1 Introduction

Partial least squares (PLS) (Wold, Sjöström, & Eriksson, 2001) is a family of techniques originally devised for modeling two sets of observed variables, which we shall call input and output components, by means of some

Neural Computation **25, 3318–3339** (2013)
doi:10.1162/NECO_a_00524

(typically low-dimensional) set of latent or unobserved components. This model can also be extended to deal with more than two sets of components (Wangen & Kowalsky, 1989). It is commonly used for regression but is also applicable to classification (Barker & Rayens, 2003). In this letter, we focus on the regression paradigm. Latent components are generated by some linear transformation of the input components, while the output components are assumed to be generated by some linear transformation of the latent components.

The difference between PLS and related techniques lies in how the latent components are estimated (Hastie, Tibshirani, & Friedman, 2008; Rosipal & Krämer, 2006). Unlike PLS, principal components regression (PCR), for example, does not consider the output when constructing the latent components. Also, PLS differs from canonical correlation analysis (CCA) in that CCA treats the input and output spaces symmetrically (Hardoon, Szedmak, & Shawe-Taylor, 2004). A complete comparison between PLS, PCR and classical shrinkage regression from a statistical perspective is given by Frank and Friedman (1993), and further insight into the shrinkage properties of PLS can be found, for instance, by Goutis (1996). There exist Bayesian formulations of some latent component models in the literature, such as PCA (Bishop, 1998; Nakajima, Sugiyama, & Babacan, 2011), CCA (Fujiwara, Miyawaki, & Kamitani, 2009; Virtanen, Klami, & Kaski, 2011; Wang, 2007) and mixtures of factor analyzers (Beal, 2003; Ghahramani & Beal, 2000). To our knowledge, however, a Bayesian version of PLS has not yet been proposed.

Different varieties of PLS regression arise by the way they extract latent components (Rosipal & Krämer, 2006). In its classic form, PLS aims to maximize the covariance among the latent components, which are constrained to be orthogonal, using the nonlinear iterative partial least squares (NIPALS) algorithm (Wold, 1975). This is more an algorithmic than a traditional statistical approach, and, hence, the analysis of its properties is less obvious. A more rigorous approach (from a statistical perspective) is taken by de Jong (1993), who directly formulates the latent space as a linear projection of the input space and solves the resulting optimization problem by the so-called SIMPLS algorithm. The SIMPLS algorithm is equivalent to NIPALS only when the output space is unidimensional. Sparsifying accounts of PLS are proposed by van Gerven, Chao, and Heskes (2012) and Chun and Keleş (2010). A kernelized approach has been introduced by Lindgren, Geladi, and Wold (1993) and Rosipal and Trejo (2001).

The main goal of this letter is to develop a Bayesian approach for PLS regression. We use variational inference (Jaakkola, 2001) for estimating the parameters. Let $\boldsymbol{X}$ be an $N \times p$ input matrix and $\boldsymbol{Y}$ be an $N \times q$ output matrix, with elements x_{ni} and y_{nj} and rows $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$. Assuming centered data, we follow the definition of PLS given by

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{P} + \boldsymbol{\epsilon}_{\boldsymbol{Z}}, \quad \boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{Q} + \boldsymbol{\epsilon}_{\boldsymbol{Y}},$$

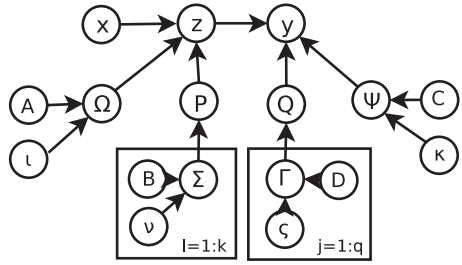


Figure 1: Bayesian hierarchy of the proposed Bayesian PLS model, where $\boldsymbol{Z}$ lies in the latent space.

where $\boldsymbol{P}$ and $\boldsymbol{Q}$ are, respectively, $p \times k$ and $k \times q$ loading matrices, $\boldsymbol{Z}$ is the $N \times k$ latent score matrix, with elements z_{il} and rows $\boldsymbol{z}_n$, and $\boldsymbol{\epsilon}_{\boldsymbol{Z}}$ and $\boldsymbol{\epsilon}_{\boldsymbol{Y}}$ are the matrices of residuals. We use an intermediate k-dimensional latent space, k typically being lower than p and q. We consider a Bayesian hierarchy defined through several normal Wishart distributions for the latent and output variables, as well as for the loading matrices:

$$\begin{aligned} \boldsymbol{z}' &\sim \mathcal{N}(\boldsymbol{x}'\boldsymbol{P}, \boldsymbol{\Omega}), \;\; \boldsymbol{\Omega}^{-1} \sim \mathcal{W}(\boldsymbol{A}, \iota), \;\; \boldsymbol{p}_l \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_l), \;\; \boldsymbol{\Sigma}_l^{-1} \sim \mathcal{W}(\boldsymbol{B}_l, \nu_l), \\ \boldsymbol{y}' &\sim \mathcal{N}(\boldsymbol{z}'\boldsymbol{Q}, \boldsymbol{\Psi}), \;\; \boldsymbol{\Psi}^{-1} \sim \mathcal{W}(\boldsymbol{C}, \kappa), \;\; \boldsymbol{q}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_j), \;\; \boldsymbol{\Gamma}_j^{-1} \sim \mathcal{W}(\boldsymbol{D}_j, \varsigma_j), \end{aligned} \tag{1.1}$$

with $l = 1, \ldots, k$ and $j = 1, \ldots, q$, and where $\boldsymbol{p}_l$ is the lth column of $\boldsymbol{P}$ and $\boldsymbol{q}_j$ is the jth column of $\boldsymbol{Q}$. $\boldsymbol{A}, \boldsymbol{B}_l, \boldsymbol{C}, \boldsymbol{D}_j$ are the scale matrix hyperparameters of the Wishart prior distributions, and $\iota, \nu_l, \kappa, \varsigma_j$ are the corresponding degrees of freedom. (See Figure 1 for a graphical representation using plate notation.) In the remainder, we suppress the hyperparameters in our notation when it is clear from the context.

By imposing separate gaussian priors for each column of $\boldsymbol{P}$ (and $\boldsymbol{Q}$), we are allowing different input (and latent) variable couplings for each component $l = 1, \ldots, k$ (and $j = 1, \ldots, q$). An obvious simplification, which we take in the rest of the letter, is to let $\boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}$ (and $\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}$), so that information is borrowed among the latent components and responses and the number of parameters is reduced. The derivation of the general case is straightforward.

It might appear that for large p scenarios, a large number of parameters is associated with the full precision matrix $\boldsymbol{\Sigma}^{-1}$. However, as we will show, the estimation of these matrices is low rank, so that the effective number of parameters is kept reasonably low.

Note that $E[\boldsymbol{y}'|\boldsymbol{x}'] = \boldsymbol{x}'\boldsymbol{P}\boldsymbol{Q}$. Since $\boldsymbol{P}\boldsymbol{Q}$ has at most rank k, this formulation is related to reduced rank methods (Izenman, 1975) and approaches that penalize the nuclear norm of the coefficient matrix (Yuan, Ekici, Lu,

& Monteiro, 2007). There is also a connection with the multivariate group Lasso (Obozinski, Wainwright, & Jordan, 2011), which imposes an L_1/L_2-penalty on the coefficient matrix so that a common sparsity pattern is shared by all responses. However, the multivariate group Lasso does not account for correlated errors. The sparse multivariate regression with covariance estimation approach (Rothman, Levina, & Zhu, 2010), on the other hand, does consider correlation between the responses by simultaneously estimating the coefficient matrix and the (full) inverse covariance matrix of the response variables. The coefficient matrix is L_1-regularized, and then the sparsity pattern can vary for each response. Our approach is also somewhat related to the multitask feature learning problem, where each task has a different set of inputs and the goal is to find some shared structural parameterization that is beneficial for the individual tasks. For example, the method proposed by Argyriou, Evgeniou, and Pontil (2006) seeks a low-rank linear transformation such that the outputs are encouraged to share a common input sparsity pattern. The method introduced by Ando and Zhang (2005) is formulated so that unlabeled data can be used for learning common underlying predictive functional structures. However, these approaches do not build on a generative, model and it is not possible to express a Bayesian formulation that leads to the same estimator.

The rest of the letter is organized as follows. Section 2 introduces the variational approximation in the basic setting. Section 3 describes how to achieve a sparse solution. Section 4 proposes an improved model that aims to estimate the optimal number of latent components and increase the accuracy. Section 5 presents a simulation study with comparisons to other methods. Section 6 provides some results for real neural signal decoding. Finally, section 7 provides conclusions and directions for future work.

2 Variational Parameter Inference

We are interested in the posterior distribution $Pr(\boldsymbol{P}, \boldsymbol{Q} \mid \boldsymbol{X}, \boldsymbol{Y})$, given by

$$\int P(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{Z}, \boldsymbol{\Omega}^{-1}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{\Psi}^{-1}, \boldsymbol{\Gamma}^{-1} \mid \boldsymbol{X}, \boldsymbol{Y})\, d\boldsymbol{Z}\, d\boldsymbol{\Omega}^{-1}\, d\boldsymbol{\Sigma}^{-1}\, d\boldsymbol{\Psi}^{-1}\, d\boldsymbol{\Gamma}^{-1}.$$

For computational reasons, we approximate the posterior distribution of the parameters given $\boldsymbol{Y}$ by a variational distribution with the following factorization:

$$\begin{aligned} &P(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{Z}, \boldsymbol{\Omega}^{-1}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{\Psi}^{-1}, \boldsymbol{\Gamma}^{-1} \mid \boldsymbol{X}, \boldsymbol{Y}) \\ &\approx F(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{Z}, \boldsymbol{\Omega}^{-1}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{\Psi}^{-1}, \boldsymbol{\Gamma}^{-1}) = F(\boldsymbol{Z})F(\boldsymbol{P}, \boldsymbol{\Omega}^{-1}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{Q}, \boldsymbol{\Psi}^{-1}, \boldsymbol{\Gamma}^{-1}). \end{aligned}$$

The variational approximation automatically (i.e., without the need for further assumptions) factorizes into $F(\boldsymbol{Z})F(\boldsymbol{P}, \boldsymbol{\Omega}^{-1}, \boldsymbol{\Sigma}^{-1})F(\boldsymbol{Q}, \boldsymbol{\Psi}^{-1}, \boldsymbol{\Gamma}^{-1})$.

This can be easily verified by inspecting the functional $F(\boldsymbol{P}, \boldsymbol{\Omega}^{-1}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{Q}, \boldsymbol{\Psi}^{-1}, \boldsymbol{\Gamma}^{-1})$, defined as the log of the joint distribution when we take the expectation with respect to $\boldsymbol{Z}$ (Beal, 2003). Since $\boldsymbol{Z}$ separates $\boldsymbol{P}, \boldsymbol{\Omega}^{-1}, \boldsymbol{\Sigma}^{-1}$ from $\boldsymbol{Q}, \boldsymbol{\Psi}^{-1}, \boldsymbol{\Gamma}^{-1}$ in the Bayesian hierarchy, the resulting expression for $F(\boldsymbol{P}, \boldsymbol{\Omega}^{-1}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{Q}, \boldsymbol{\Psi}^{-1}, \boldsymbol{\Gamma}^{-1})$ does not have interaction terms between the two groups of variables.

Also on computational grounds, we assume $F(\boldsymbol{P}, \boldsymbol{\Omega}^{-1}, \boldsymbol{\Sigma}^{-1}) = F(\boldsymbol{P})$ $F(\boldsymbol{\Omega}^{-1}, \boldsymbol{\Sigma}^{-1})$ and, analogously, $F(\boldsymbol{Q}, \boldsymbol{\Psi}^{-1}, \boldsymbol{\Gamma}^{-1}) = F(\boldsymbol{Q})F(\boldsymbol{\Psi}^{-1}, \boldsymbol{\Gamma}^{-1})$. From this, we have an automatic factorization between $\boldsymbol{\Omega}^{-1}$ and $\boldsymbol{\Sigma}^{-1}$ ($\boldsymbol{\Psi}^{-1}$ and $\boldsymbol{\Gamma}^{-1}$). Finally, $F(\boldsymbol{Z})$ automatically factorizes into $\prod_{n=1}^{N} F(\boldsymbol{z}_n)$, so that up to a constant, we have

$$\begin{aligned} F(\boldsymbol{z}_n) &= E_{\boldsymbol{P},\boldsymbol{Q},\boldsymbol{\Omega}^{-1},\boldsymbol{\Psi}^{-1}}[\log P(\boldsymbol{z}_n|\boldsymbol{X}, \boldsymbol{P}, \boldsymbol{\Omega}^{-1}) + \log P(\boldsymbol{y}|\boldsymbol{z}_n, \boldsymbol{Q}, \boldsymbol{\Psi}^{-1})] \\ &= -\frac{1}{2}\boldsymbol{z}_n'(E[\boldsymbol{\Omega}^{-1}]+E[\boldsymbol{Q}\boldsymbol{\Psi}^{-1}\boldsymbol{Q}'])\boldsymbol{z}_n+(\boldsymbol{x}_n'\boldsymbol{\mu}_{\boldsymbol{P}}E[\boldsymbol{\Omega}^{-1}]+\boldsymbol{y}_n'E[\boldsymbol{\Psi}^{-1}]\boldsymbol{\mu}_{\boldsymbol{Q}})\boldsymbol{z}_n, \end{aligned}$$

where $\boldsymbol{\mu}_{\boldsymbol{P}}$ and $\boldsymbol{\mu}_{\boldsymbol{Q}}$ are the expectations of, respectively, $\boldsymbol{P}$ and $\boldsymbol{Q}$. Expectations are with regard to the variational distribution. Completing the square, we have

$$F(\boldsymbol{z}_n) = \mathcal{N}(\boldsymbol{z}_n; \boldsymbol{\mu}_{\boldsymbol{z}_n}, \boldsymbol{S}_{\boldsymbol{z}_n}) \tag{2.1}$$

with $\boldsymbol{S}_{\boldsymbol{z}_n} = (E[\boldsymbol{\Omega}^{-1}]+E[\boldsymbol{Q}\boldsymbol{\Psi}^{-1}\boldsymbol{Q}'])^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{z}_n} = \boldsymbol{S}_{\boldsymbol{z}_n}(E[\boldsymbol{\Omega}^{-1}]\boldsymbol{\mu}_{\boldsymbol{P}}'\boldsymbol{x}_n + \boldsymbol{\mu}_{\boldsymbol{Q}}E[\boldsymbol{\Psi}^{-1}]$ $\boldsymbol{y}_n)$. We can compute

$$E[\boldsymbol{Q}\boldsymbol{\Psi}^{-1}\boldsymbol{Q}'] = \boldsymbol{\mu}_{\boldsymbol{Q}}E[\boldsymbol{\Psi}^{-1}]\boldsymbol{\mu}_{\boldsymbol{Q}}' + \sum_{j_1=1}^{q}\sum_{j_2=1}^{q} E[\Psi_{j_1 j_2}^{-1}]\boldsymbol{S}_{\boldsymbol{Q}_{j_1 j_2}},$$

where $\boldsymbol{S}_{\boldsymbol{Q}_{j_1 j_2}}$ denotes the $k \times k$ cross-covariance matrix relative to the j_1th and j_2th columns of the loading matrix $\boldsymbol{Q}$.

For $\boldsymbol{\Omega}^{-1}$, we have, up to a constant,

$$\begin{aligned} \log F(\boldsymbol{\Omega}^{-1}) &= E_{\boldsymbol{Z},\boldsymbol{P}}[\log P(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{P}, \boldsymbol{\Omega}^{-1}) + \log P(\boldsymbol{\Omega}^{-1})] \\ &= \frac{\iota - k - N - 1}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\mathrm{Tr}((E[\boldsymbol{Z}'\boldsymbol{Z}] + E[\boldsymbol{P}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{P}] \\ &\quad -\boldsymbol{\mu}_{\boldsymbol{Z}}'\boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{P}} - \boldsymbol{\mu}_{\boldsymbol{P}}'\boldsymbol{X}'\boldsymbol{\mu}_{\boldsymbol{Z}} + \boldsymbol{A}^{-1})\,\boldsymbol{\Omega}^{-1}), \end{aligned}$$

where we have used standard properties of the trace operator. Here, we can identify a Wishart distribution,

$$F(\boldsymbol{\Omega}^{-1}) = \mathcal{W}(\boldsymbol{\Omega}^{-1}; \tilde{\mathbf{A}}^{-1}, \tilde{\iota}), \tag{2.2}$$

with $\tilde{\iota} = \iota + N$ and $\tilde{\mathbf{A}}^{-1} = (E[\boldsymbol{Z}'\boldsymbol{Z}] + E[\boldsymbol{P}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{P}] - \boldsymbol{\mu}_{\boldsymbol{Z}}'\boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{P}} - \boldsymbol{\mu}_{\boldsymbol{P}}'\boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{Z}} + \boldsymbol{A}^{-1})^{-1}$, where $E[\boldsymbol{Z}'\boldsymbol{Z}] = \sum_{n=1}^{N} E[\boldsymbol{z}_n\boldsymbol{z}_n'] = \sum_{n=1}^{N}(\boldsymbol{S}_{\boldsymbol{z}_n} + \boldsymbol{\mu}_{\boldsymbol{z}_n}\boldsymbol{\mu}_{\boldsymbol{z}_n}')$.

If we set $\boldsymbol{\Omega}^{-1}$ to be diagonal, then the variational distribution $F(\boldsymbol{P})$ factorizes over columns and we get a gamma distribution for each element Ω_{ll}^{-1}:

$$F(\Omega_{ll}^{-1}) = \mathcal{G}(\Omega_{ll}^{-1}; \tilde{\iota}, \tilde{A}_{ll}^{-1}) \tag{2.3}$$

with $\tilde{\iota} = \iota + \frac{N}{2}$ and $\tilde{A}_{ll}^{-1} = \frac{1}{2}(E[\boldsymbol{Z}_{\cdot l}\boldsymbol{Z}_{\cdot l}'] + E[\boldsymbol{p}_l'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{p}_l] - 2\,\boldsymbol{\mu}_{\boldsymbol{Z}_{\cdot l}}'\boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{p}_l}) + A_{ll}^{-1}$, where $\boldsymbol{Z}_{\cdot l}$ denotes the lth column of $\boldsymbol{Z}$.

If we do not factorize $F(\boldsymbol{P})$, that is, if $\boldsymbol{\Omega}^{-1}$ is not chosen to be diagonal, then we have, up to a constant,

$$\begin{aligned}
&\log F(\boldsymbol{P}) \\
&\quad = E_{\boldsymbol{Z},\boldsymbol{\Sigma}^{-1},\boldsymbol{\Omega}^{-1}}\left[\log P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{P},\boldsymbol{\Omega}^{-1}) + \sum_{l=1}^{k}\log P(\boldsymbol{p}_l|\boldsymbol{\Sigma})\right] \\
&\quad = -\sum_{n=1}^{N}\left(\frac{1}{2}\boldsymbol{x}_n'\boldsymbol{P}E[\boldsymbol{\Omega}^{-1}]\boldsymbol{P}'\boldsymbol{x}_n - \boldsymbol{x}_n'\boldsymbol{P}E[\boldsymbol{\Omega}^{-1}]\boldsymbol{\mu}_{\boldsymbol{z}_n}\right) - \frac{1}{2}\sum_{l=1}^{k}\boldsymbol{p}_l'E[\boldsymbol{\Sigma}^{-1}]\boldsymbol{p}_l.
\end{aligned}$$

We define $\tilde{\boldsymbol{p}}$ as the concatenation of the rows of $\boldsymbol{P}$ and $\tilde{\boldsymbol{p}}^*$ as the concatenation of the rows of the $p \times k$ least-squares solution $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\mu}_{\boldsymbol{Z}}$, so that after some algebra, we can identify a pk-dimensional gaussian distribution,

$$F(\boldsymbol{P}) = \mathcal{N}(\boldsymbol{P}; \boldsymbol{\mu}_{\boldsymbol{P}}, \boldsymbol{S}_{\boldsymbol{P}}) \tag{2.4}$$

with $\boldsymbol{S}_{\tilde{\boldsymbol{p}}} = (E[\boldsymbol{\Sigma}^{-1}] \otimes \boldsymbol{I}_k + \boldsymbol{X}'\boldsymbol{X} \otimes E[\boldsymbol{\Omega}^{-1}])^{-1}$ and $\boldsymbol{\mu}_{\tilde{\boldsymbol{p}}} = \boldsymbol{S}_{\tilde{\boldsymbol{p}}}\big(\boldsymbol{X}'\boldsymbol{X} \otimes E[\boldsymbol{\Omega}^{-1}]\big)\tilde{\boldsymbol{p}}^*$, where $\boldsymbol{I}_k$ is the $k \times k$ identity matrix and $\otimes$ denotes the Kronecker product. From this expression, we can reconstruct $\boldsymbol{\mu}_{\boldsymbol{P}}$ and $\boldsymbol{S}_{\boldsymbol{P}}$.

When $\boldsymbol{\Omega}^{-1}$ is diagonal, we can simplify $F(\boldsymbol{P}) = \prod_{l=1}^{k} F(\boldsymbol{p}_l)$. For each factor, we have

$$F(\boldsymbol{p}_l) = \mathcal{N}(\boldsymbol{p}_l; \boldsymbol{\mu}_{\boldsymbol{p}_l}, \boldsymbol{S}_{\boldsymbol{p}_l}) \tag{2.5}$$

with $\boldsymbol{S}_{\boldsymbol{p}_l} = (E[\boldsymbol{\Sigma}^{-1}] + E[\Omega_{ll}^{-1}]\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{p}_l} = E[\Omega_{ll}^{-1}]\boldsymbol{S}_{\boldsymbol{p}_l}\boldsymbol{X}'\boldsymbol{\mu}_{\boldsymbol{Z}_{\cdot l}}$.

For $\boldsymbol{\Sigma}^{-1}$, we have, up to a constant,

$$\begin{aligned}
\log F(\boldsymbol{\Sigma}^{-1}) &= E_{\boldsymbol{p}_l}\left[\log P(\boldsymbol{\Sigma}^{-1}) + \sum_{l=1}^{k}\log P(\boldsymbol{p}_l|\boldsymbol{\Sigma}^{-1})\right] \\
&= \frac{\nu_l - p - k - 1}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\mathrm{Tr}((E[\boldsymbol{P}\boldsymbol{P}'] + \boldsymbol{B}_l^{-1})\boldsymbol{\Sigma}^{-1}),
\end{aligned}$$

where we can identify a Wishart distribution:

$$F(\boldsymbol{\Sigma}^{-1}) = \mathcal{W}(\boldsymbol{\Sigma}^{-1}; \tilde{\boldsymbol{B}}^{-1}, \tilde{\boldsymbol{\nu}}), \tag{2.6}$$

with $\tilde{\boldsymbol{B}}^{-1} = (E[\boldsymbol{P}\boldsymbol{P}'] + \boldsymbol{B}^{-1})^{-1}$ and $\tilde{\nu} = \nu + k$.

Note that the matrix $E[\boldsymbol{P}\boldsymbol{P}']$ is not full rank as far as $p > k$, which is typically the case. It has, in fact, rank k. Then the effective number of parameters of this matrix is not $p(p-1)/2$, but, at most, $pk + 1 - k(k-1)/2$. When p is high relative to N, it becomes necessary to borrow information between the components $l = 1, \ldots, k$, suggesting the choice $\boldsymbol{\Sigma}_l^{-1} = \boldsymbol{\Sigma}^{-1}$.

Calculations for $\boldsymbol{Q}$ and dependent distributions are similar to those of $\boldsymbol{P}$ and are given in appendix A. Brown & Zidek (1980) theoretically showed that an adaptive joint estimation dominates an independent estimation for each of the outputs separately when the number of inputs is considerably larger than the number of outputs, but this domination breaks down when the number of outputs approaches the number of inputs. In our situation, when estimating $\boldsymbol{Q}$, the number of outputs q typically even exceeds the number of hidden units k. This suggests that the factorization $F(\boldsymbol{Q}) = \prod_{j=1}^{q} F(\boldsymbol{q}_j)$, mimicking independent estimation, is the sensible choice for $q > k$, which is often the case.

In short, the proposed approach proceeds as follows:

1. Initialize $\boldsymbol{Z}$ to the k first principal components of $\boldsymbol{Y}$.
2. Compute the distributions of $\boldsymbol{\Omega}^{-1}$, $\boldsymbol{P}$ and $\boldsymbol{\Sigma}^{-1}$ using equations 2.2, 2.4, and 2.6.
3. Compute the distributions of $\boldsymbol{\Psi}^{-1}$, $\boldsymbol{Q}$ and $\boldsymbol{\Gamma}^{-1}$ using equations A.1, A.3, and A.5.
4. Compute the distribution of $\boldsymbol{Z}$ using equation 2.1.
5. Repeat steps 2 to 4 until convergence.

This grouping of the updates is motivated by the structure of the Bayesian hierarchy and the variational factorization in equation 2.1. A variant of the basic scheme, by assuming diagonality of $\boldsymbol{\Omega}^{-1}$ and $\boldsymbol{\Psi}^{-1}$, arises by substituting equation 2.2 by 2.3, 2.4 by 2.5, A.1 by A.2, and A.3 by A.4. A variational lower bound of the evidence is given in appendix B.

3 Sparsity in P and Q

For achieving sparsity on the input variables, we may impose a group-sparsifying prior on $\boldsymbol{P}$, so that the groups are the k-dimensional rows of $\boldsymbol{P}$. By setting an individual regularization parameter on each group and integrating out $\boldsymbol{P}$, the maximum likelihood value of such regularization parameters will effectively discard some groups. This is an example of groupwise automatic relevance determination (see, for example, Virtanen et al., 2011).

To achieve this objective, we set priors

$$\boldsymbol{P}_{i\cdot} \sim \mathcal{N}(\mathbf{0}, \sigma_i^2 \boldsymbol{I}_k),$$

where $\boldsymbol{P}_{i\cdot}$ is the ith row of $\boldsymbol{P}$. This way, we will effectively drop the useless inputs (rows of $\boldsymbol{P}$). We define the precisions σ_i^{-2} to be gamma distributed. The variational approximation of the posterior of $\boldsymbol{p}_l$ becomes a gaussian distribution with parameters $\boldsymbol{S}_{\boldsymbol{p}_l} = (\mathrm{diag}(E[\boldsymbol{\sigma}^{-2}]) + E[\Omega_{ll}^{-1}]\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{p}_l} = E[\Omega_{ll}^{-1}]\boldsymbol{S}_{\boldsymbol{p}_l}\boldsymbol{X}'\boldsymbol{\mu}_{\boldsymbol{Z}_{\cdot l}}$.

The derivation for nonfactorized $\boldsymbol{P}$ is straightforward. For σ_i^{-2}, we have

$$E[\sigma_i^{-2}] = \frac{2\nu + k}{E[\boldsymbol{P}_{i\cdot}'\boldsymbol{P}_{i\cdot}] + 2B_{ii}^{-1}} = \frac{2\nu + k}{\sum_{l=1}^{k}(S_{\boldsymbol{P}_{il,il}} + \mu_{\boldsymbol{P}_{il}}^2) + 2B_{ii}^{-1}}.$$

Also, we impose a similar groupwise prior on $\boldsymbol{Q}$:

$$\boldsymbol{Q}_{l\cdot} \sim \mathcal{N}(\mathbf{0}, \gamma_l^2 \boldsymbol{I}_q),$$

with the precision γ_l^{-2} being gamma distributed. The idea is to obtain a data-driven estimation of the importance of each latent component when estimating $\boldsymbol{Q}$. The variational approximation of $\boldsymbol{q}_j$ is a gaussian distribution with parameters $\boldsymbol{S}_{\boldsymbol{q}_j} = (\mathrm{diag}(E[\boldsymbol{\gamma}^{-2}]) + E[\Psi_{jj}^{-1}]E[\boldsymbol{Z}'\boldsymbol{Z}])^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{q}_j} = E[\Psi_{jj}^{-1}]\boldsymbol{S}_{\boldsymbol{q}_j}\boldsymbol{\mu}_{\boldsymbol{Z}}'\boldsymbol{Y}_{\cdot j}$.

Again, we follow a variational approximation to obtain

$$E[\gamma_l^{-2}] = \frac{2\varsigma + q}{E[\boldsymbol{Q}_{l\cdot}'\boldsymbol{Q}_{l\cdot}] + 2D_{ll}^{-1}} = \frac{2\varsigma + q}{\sum_{j=1}^{q}(S_{\boldsymbol{Q}_{lj,lj}} + \mu_{\boldsymbol{Q}_{lj}}^2) + 2D_{ll}^{-1}}.$$

Then a value σ_i^{-2} (γ_l^{-2}) close to zero means that the corresponding input (latent) variable can be considered irrelevant.

4 Adaptive P and Automatic Selection of k

The proposed estimation of $\boldsymbol{P}$ disregards $\boldsymbol{Y}$ and uses only the current state of $\boldsymbol{Z}$. Put differently, equal attention is paid to all latent components when estimating $\boldsymbol{P}$, no matter the contribution of each latent component to the prediction of $\boldsymbol{Y}$. A possible improvement would be to focus on those latent components that are more useful for modeling $\boldsymbol{Y}$, regularizing more the latent components that are relatively useless—those whose coefficients in $\boldsymbol{Q}$ are lower.

Here, we impose groupwise priors on $\boldsymbol{P}$ for both rows and columns so as to achieve two goals. First, we provide a more adaptive estimation of

$\boldsymbol{P}$, which we hope will reduce the bias. Second, starting with a reasonably high value of k, we will be better able to obtain an estimate of the adequate dimension for the latent space. Hence, the purpose of this section is to improve the model performance for a given number of latent components. A proper model selection scheme, which would compare the model evidence (see appendix B) for different values of k, would be an alternative (or complementary) way to go.

We consider the variational approximation with $\boldsymbol{P}$ and $\boldsymbol{Q}$ factorized over columns. We let $\boldsymbol{\gamma}^{-2}$ also influence $\boldsymbol{P}$ so that it indeed controls the latent space for both projection matrices. We balance the regularization effect of $\boldsymbol{\gamma}^{-2}$ with an additional variable ϕ^{-2}, so that the relative contributions of $\boldsymbol{\sigma}^{-2}$ and $\boldsymbol{\gamma}^{-2}$ are adaptively estimated from data. We choose ϕ^{-2} to be gamma distributed for keeping conjugacy. Specifically, we propose

$$P_{il} \sim \mathcal{N}(\mathbf{0}, (\sigma_i^{-2} + \phi^{-2}\gamma_l^{-2})^{-1}), \qquad \phi^{-2} \sim \mathcal{G}(e, \varphi).$$

The variational distribution of $\boldsymbol{p}_l$ is a gaussian distribution whose parameters are $\boldsymbol{S}_{\boldsymbol{p}_l} = (\mathrm{diag}(E[\boldsymbol{\sigma}^{-2}]) + E[\phi^{-2}]E[\gamma_l^{-2}]\boldsymbol{I}_p + E[\Omega_{ll}^{-1}]\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{p}_l} = E[\Omega_{ll}^{-1}]\boldsymbol{S}_{\boldsymbol{p}_l}\boldsymbol{X}'\boldsymbol{\mu}_{\boldsymbol{Z}_{.l}}$.

For $\boldsymbol{\gamma}^{-2}$, we have

$$E[\gamma_l^{-2}] = \frac{2\varsigma + p + q}{\phi^{-2}\sum_{i=1}^{p}(S_{\boldsymbol{P}_{il,il}} + \mu^2_{\boldsymbol{P}_{il}}) + \sum_{j=1}^{q}(S_{\boldsymbol{Q}_{lj,lj}} + \mu^2_{\boldsymbol{Q}_{lj}}) + 2d_l^{-1}},$$

and, for ϕ^{-2},

$$E[\phi^{-2}] = \frac{2\varphi + pk}{\sum_{i=1}^{p}\sum_{l=1}^{k}\gamma_l^{-2}(S_{\boldsymbol{P}_{il,il}} + \mu^2_{\boldsymbol{P}_{il}}) + 2e^{-1}}.$$

Then, by automatic relevance determination, the number of coordinates of $\boldsymbol{\gamma}^{-2}$ significantly greater than zero at convergence is an estimation of the optimal number of latent variables k.

5 Synthetic Experiments

In order to test in practice the performance of the algorithms, we now present a synthetic simulation study where we compare the sparse approach proposed in section 3 (SPLS for short) with other methods. For simplicity, we do not present results for the nonfactorized estimation of $\boldsymbol{P}$ and $\boldsymbol{Q}$ or for full matrices $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Gamma}^{-1}$.

We have tested some alternative PLS methods: the NIPALS and SIMPLS algorithms and the frequentist sparse PLS from Chun and Keleş (2010) (sSIMPLS). All the PLS methods have been provided with four latent

components. Also, we have tested a number of non-PLS univariate and multivariate regression techniques. Univariate methods are separately applied for each response; these include ordinary least squares regression (OLS), ridge regression, and the Lasso (Hastie et al., 2008). From the multivariate side, we have tested the multivariate group lasso (MGL) (Obozinski et al., 2011) and the sparse multivariate regression with covariance estimation approach (MRCE) (Rothman et al., 2010). The regularization parameters for the non-Bayesian methods have been chosen by cross-validation.

The design of the simulation study is as follows. In all experiments, the number of input variables is $p = 50$, and the number of responses is $q = 8$. We have covered several different situations, varying the number of hidden components, which we denote as k_0, and the amount of training data ($N = 100, 500$). We have set $k_0 = 1, 2, 4, 8$. Within each situation, we generated 100 random replications. For each of them, we have sampled 1000 testing data points. Then, for each situation and each model, reports are shown over $100 \times 1000 \times 8 = 800{,}000$ residuals.

For each random replication, the input matrix was generated according to a multivariate gaussian distribution with zero mean and covariance matrix $\boldsymbol{M}$, whose elements were set as $M_{i_1 i_2} = r_1^{|i_1 - i_2|}$, and r_1 was sampled from the uniform distribution with support $[0, 1]$. Hence, depending on r_1, the degree of collinearity in the input matrix is different among the replications. Of course, the testing input matrix was sampled using the same covariance matrix $\boldsymbol{M}$.

When $k_0 < 8$, the latent components were generated as $\boldsymbol{Z} = \boldsymbol{XP} + \boldsymbol{\epsilon_Z}$. Sparsity was enforced by setting each row of $\boldsymbol{P}$ to zero with probability 0.8 (ensuring at least two relevant input variables). The rest of the rows were sampled from a standard normal distribution. The noise $\boldsymbol{\epsilon_Z}$ was generated from a gaussian distribution with zero mean and diagonal covariance matrix $\boldsymbol{\Omega}$, with $\Omega_{ll} = r_2 \operatorname{sd}(\boldsymbol{X}\boldsymbol{p}_l)$, where $\operatorname{sd}(\cdot)$ denotes the sample standard deviation and the value r_2 was sampled from the uniform distribution with support $[0.01, 0.1]$ (separately for each each diagonal element).

We generated the responses as $\boldsymbol{Y} = \boldsymbol{ZQ} + \boldsymbol{\epsilon_Y}$. We did not consider sparsity in $\boldsymbol{Q}$, whose elements were sampled from a standard normal distribution. The noise $\boldsymbol{\epsilon_Y}$ was sampled from a gaussian distribution with zero mean and diagonal covariance matrix $\boldsymbol{\Psi}$, with diagonal elements $\Psi_{jj} = r_3 \operatorname{sd}(\boldsymbol{Z}\boldsymbol{q}_j)$, where r_3 was sampled from the uniform distribution with support $[0.25, 0.5]$.

When $k_0 = 8$, the response was computed as $\boldsymbol{Y} = \boldsymbol{XF} + \boldsymbol{\epsilon_Y}$, where $\boldsymbol{F}$ has rank q. Since $q > k$, $\boldsymbol{F}$ has a higher number of effective parameters than $\boldsymbol{PQ}$ in the factorized ($k_0 < 8$) case. $\boldsymbol{F}$ was sampled from a normal standard distribution, considering sparsity as before.

Figures 2, 3, 4, and 5 show box plots with the performance of the different methods for all situations. Results are in terms of the explained variance R^2. In short, these graphs illustrate how the methods behave for different

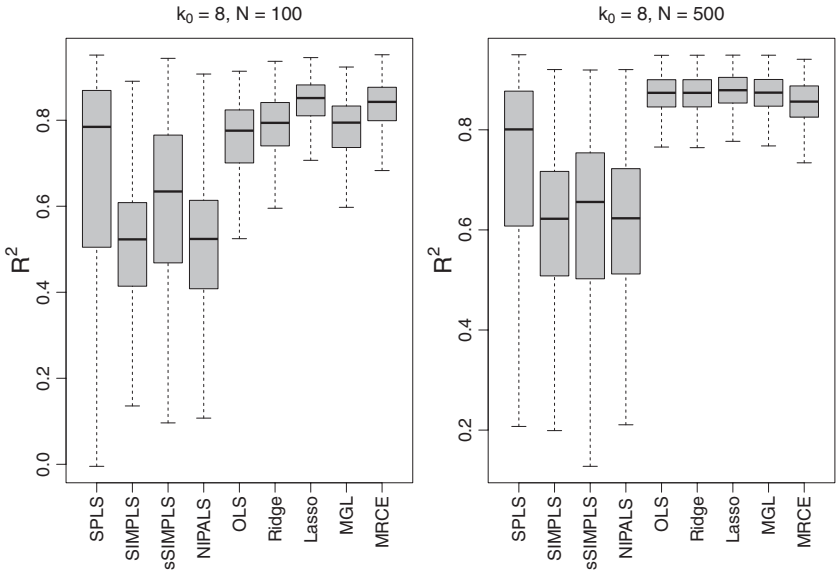


Figure 2: Box plot of the R^2 values for $k_0 = 8$ and $N = 100, 500$.

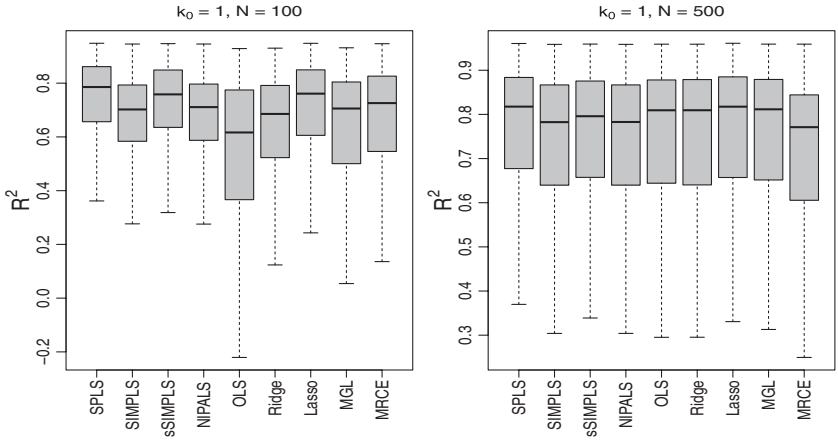


Figure 3: Box plot of the R^2 values for $k_0 = 1$ and $N = 100, 500$.

complexities of the true response function and amounts of training input data.

When $k_0 = 8$, the response function does not factorize and has the highest number of parameters. In this case, the PLS methods have fewer parameters than the actual true response function and tend to underfit. This is in

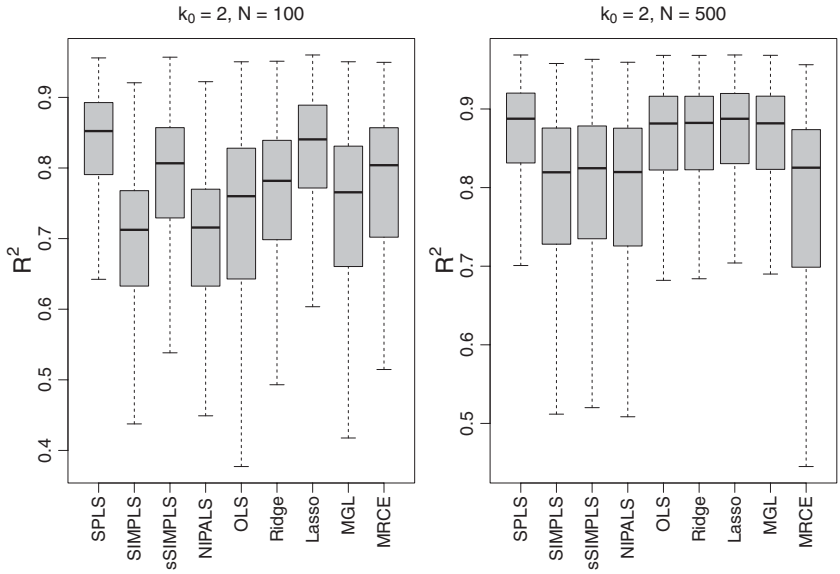


Figure 4: Box plot of the R^2 values for $k_0 = 2$ and $N = 100, 500$.

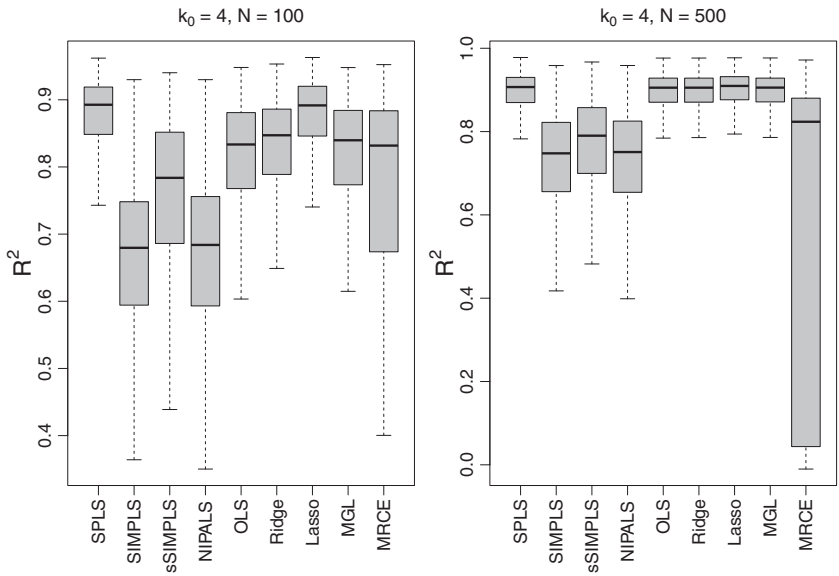


Figure 5: Box plot of the R^2 values for $k_0 = 4$ and $N = 100, 500$.

particular the case for $N = 500$, when there are sufficient data for a reliable estimation of all the parameters. Interestingly, SPLS is the only PLS method that does not perform much worse than ridge regression, the Lasso, MGL, and MRCE for $N = 100$ and also for $N = 500$.

The $k_0 = 1$ case is the opposite extreme. Whereas the full-rank methods aim to estimate $pq = 400$ parameters, the PLS methods, which are set to use four latent components, estimate $4p + 4q = 232$ parameters. Still, this is much more than the true number of parameters ($p + q = 58$). For this reason, the PLS methods are not much better than the others. Indeed, SPLS and the Lasso appear to be the more effective in controlling the complexity of the model. Note that the differences between the methods are very subtle for $N = 500$.

When $k_0 = 2$, the true response function has 116 parameters. In this case, for $N = 100$, SPLS and the Lasso clearly outperform the other methods. SPLS is slightly better than the Lasso. Surprisingly, SIMPLS and NIPALS do not perform better than OLS. For $N = 500$, SPLS, OLS, ridge regression, the Lasso, and MGL are almost indistinguishable and better than the others.

Finally, when $k_0 = 4$, the true number of parameters (232) matches that of the PLS methods. Surprisingly, for $N = 100$, SIMPLS, sSIMPLS, and NIPALS are again less accurate than any of the univariate regression methods, MGL and MRCE. On the contrary, SPLS, followed by the Lasso, are the better methods. For $N = 500$, SPLS, the univariate regression methods (including OLS) and MGL have the same performance, which is not far from the optimal Bayes error.

It is noticeable that SPLS is the only PLS method that works as well as the Lasso, and even beats it in some situations ($N = 100$ and $k_0 = 1, 2$). The good performance of the univariate regression techniques (in particular, ridge regression and the Lasso) is very likely because $\mathbf{\Psi}$ is diagonal, even when there is a coupling due to $\mathbf{\Omega}$. The poor performance of MRCE probably comes from the estimation of a full inverse covariance matrix of the response variables.

In these experiments, the performance of APLS (not shown) is not very different from that of SPLS. This is not surprising, because the synthetic data sets were generated according to equations 1.1, which correspond to the SPLS model. In the next section, we shall see an example where the adaptive version is clearly better.

6 Electrocorticogram Data Decoding

In this section we describe some experimental results on a neuroscientific data set. In particular, we aim to decode the motor output from electrocorticogram (ECoG) signals collected in monkeys. ECoG signals were recorded at a sampling rate of 1 kHz per channel, for 32 electrodes implanted in the right hemisphere, and filtered. Afterward, the signals were bandpass-filtered from 0.3 to 500 Hz. A time-frequency representation of

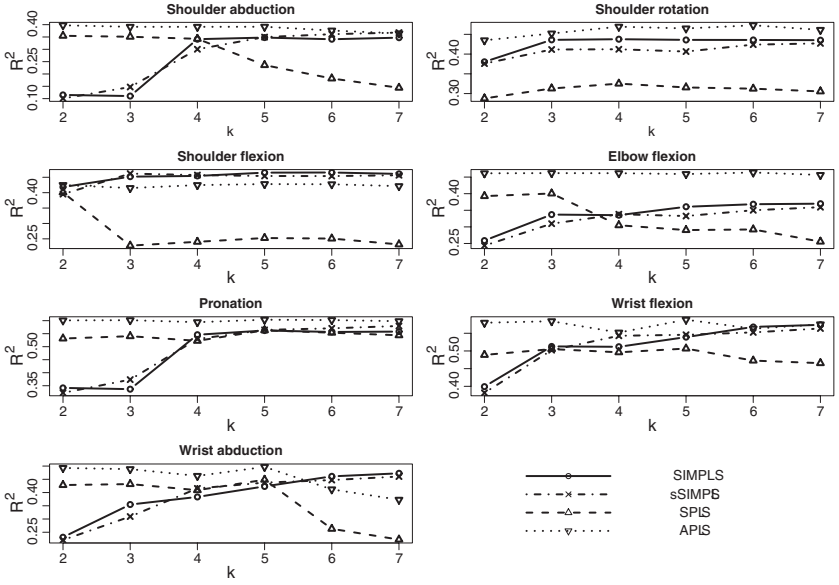
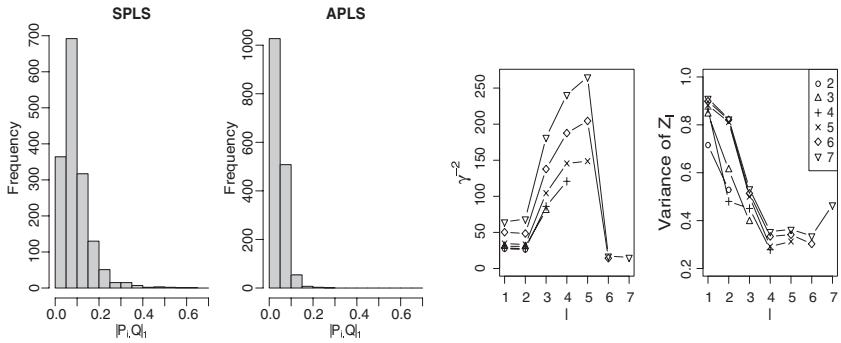


Figure 6: R^2 value for each response for a different number of latent components $k = 2, \ldots, 7$.

the ECoG signals containing $p = 1600$ variables (32 electrodes, 10 frequency bins, and 5 time lags) was used as the input for the algorithms. The monkey's movements were captured at a sampling rate of 120 Hz. Seven degrees of freedom of the monkey's movements ($q = 7$) are to be modeled and predicted. From the available data, we have used $N = 5982$ time points for training and 5982 more for testing. More details about the data can be found in van Gerven et al. (2012).

Given the high dimensionality of the data, we have run the simplest version of the algorithms; we have factorized $\boldsymbol{P}$ and $\boldsymbol{Q}$, and $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Gamma}^{-1}$ are taken to be diagonal.

Figure 6 illustrates the performance (explained variance R^2 over testing data) of the proposed approaches compared to SIMPLS and sSIMPLS. Although not included in the plot, the results of NIPALS are almost indistinguishable from SIMPLS. It is remarkable that APLS performs better than the other methods for most of the responses and is always better than the nonadaptive algorithm. Note that APLS appears to need fewer latent components to give a reasonable estimation. Moreover, APLS is more robust to the choice of k than the other algorithms (including SPLS). Only for the wrist abduction response is the APLS accuracy clearly decreased when $k > 5$. On the other hand, SPLS performs worse in general for high values of k.



a. b.

Figure 7: (a) Histogram of values $\sum_{j=1}^{q} |\boldsymbol{P}'_{i\cdot}\boldsymbol{q}_j|$ for SPLS and APLS. (b) Vector $\boldsymbol{\gamma}^{-2}$ and vector of individual latent component variances for APLS. Each line, labeled with a different type of symbol, represents a different run with a different maximum number of latent components.

Figure 7a shows, for SPLS and APLS, a histogram with the values of $\sum_{j=1}^{q} |\boldsymbol{P}'_{i\cdot}\boldsymbol{q}_j|$ as a measure of the importance of each input variable for the output prediction. (Note that $\boldsymbol{\sigma}^{-2}$ reflects the importance of each input for predicting the latent variables and hence is not a good measure of importance of each input variable for the output prediction.) It is worth noting that APLS yields sparser solutions than SPLS in this sense.

Figure 7b shows the values of $\boldsymbol{\gamma}^{-2}$ and the variance of the latent components for six executions of APLS, each with a different number of latent components $k = 2, \ldots, 7$. Each line, labeled with a different type of symbol, corresponds to a different value k. Note that latent variables that have a high value ${\gamma_l}^{-2}$ (left graph) or exhibit a low variance (right graph) are given less importance. From the graph, we can conclude that the first two latent components are the most important for the prediction. For example, the line with the $\times$ symbol corresponds to $k = 5$ and has five components (symbols). Each component of this line corresponds to a latent component in the model. The lowest value of ${\gamma_l}^{-2}$ (or the highest variance of $\boldsymbol{Z}_l$) for this model pertains to the first two components. Hence, for this model, the two relevant components are $l = 1, 2$. For the models with $k = 6, 7$ components (whose lines have the $\diamond$ and $\triangledown$ symbols), however, the last components have the lowest values for $\boldsymbol{\gamma}^{-2}$. The accuracy in these cases is worse than the accuracy that can be obtained with lower k values (see Figure 6). In summary, for all runs, there are two predominant latent components (either the first two or the last two), which indicates that $k = 2$ is a reasonable estimate of the optimal value of latent components.

Finally, Figure 8 demonstrates the trajectories decoded by SIMPLS and APLS. We have used $k = 7$ for SIMPLS and $k = 2$ for APLS. The two

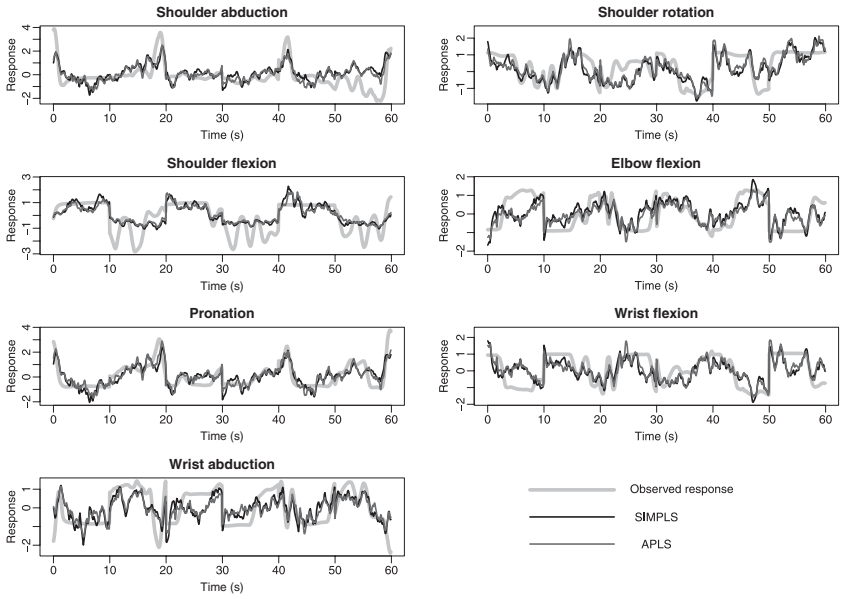


Figure 8: Observed and estimated responses in a 60 second time window, as predicted by SIMPLS and APLS. Two latent variables for APLS and seven latent variables for SIMPLS have been used.

algorithms turn out to do well. APLS is a bit smoother (less noisy) than SIMPLS. Again, the outcome for NIPALS is very similar to SIMPLS.

Table 1 illustrates the results of APLS versus OLS, ridge regression, the Lasso, MGL, and MRCE. Interestingly, the performance of APLS is higher than the other methods for all responses. Most differences are statistically significant according to a t-test. MGL does worse than ridge regression and the Lasso. MRCE, however, is also quite competitive.

7 Discussion

We have proposed a Bayesian formulation of PLS, with extensions for sparsity, adaptive modeling of $\boldsymbol{P}$, and automatic determination of k, and we empirically showed that they perform well on ECoG data decoding.

The proposed approximation relies on the Bayesian paradigm, and, hence, regularization is performed in a data-driven fashion with low risk of overfitting. An advantage is interpretability of the model: using diagonal matrices $\boldsymbol{\Sigma}^{-1} = \mathrm{diag}(\boldsymbol{\sigma}^{-2})$ and $\boldsymbol{\Gamma}^{-1} = \mathrm{diag}(\boldsymbol{\gamma}^{-2})$, automatic relevance determination provides a measure of the relevance of each input and latent component. If, in addition, we use the adaptive extension proposed in section 4, we can obtain a reasonable estimate of the optimal number k of latent

Table 1: Mean Absolute Error (and Standard Deviations) for APLS ($k = 2$), OLS, Univariate Ridge, Univariate Lasso, Multivariate Group Lasso (MGL), and Multivariate Regression with Covariance Estimation (MRCE).

	APLS	OLS	Ridge	Lasso	MGL	MRCE
SA	**0.59**($\pm$0.003)	0.92($\pm$0.004)	0.70($\pm$0.003)	0.77($\pm$0.003)	0.90($\pm$0.004)	0.60($\pm$0.003)
SF	**0.64**($\pm$0.003)	0.92($\pm$0.004)	0.72($\pm$0.003)	0.79($\pm$0.003)	0.91($\pm$0.004)	0.72($\pm$0.002)
P	**0.51**($\pm$0.002)	0.70($\pm$0.003)	0.56($\pm$0.002)	0.60($\pm$0.002)	0.70($\pm$0.003)	**0.51**($\pm$0.002)
WA	**0.57**($\pm$0.002)	0.84($\pm$0.004)	0.62($\pm$0.003)	0.74($\pm$0.003)	0.84($\pm$0.003)	0.68($\pm$0.002)
SR	**0.55**($\pm$0.002)	0.81($\pm$0.003)	0.61($\pm$0.002)	0.70($\pm$0.003)	0.80($\pm$0.003)	0.57($\pm$0.002)
EF	**0.53**($\pm$0.002)	0.84($\pm$0.004)	0.70($\pm$0.003)	0.68($\pm$0.003)	0.81($\pm$0.003)	0.66($\pm$0.002)
WF	**0.54**($\pm$0.002)	0.75($\pm$0.003)	0.58($\pm$0.002)	0.65($\pm$0.002)	0.74($\pm$0.003)	0.66($\pm$0.002)

Notes: Each row corresponds to a motor output: shoulder abduction (SA), shoulder flexion (SF), pronation (P), wrist abduction (WA), shoulder rotation (SR), elbow flexion (EF), and wrist flexion (WF). The best method is highlighted in bold.

components from $\boldsymbol{\gamma}^{-2}$. Unlike other PLS formulations, the adaptive model appears to be robust to the choice of k.

For automatically selecting k, we can run the algorithm with several values of k and then select the one that reaches the highest model evidence (see appendix B). A more sophisticated solution is to move toward a nonparametric method, where a proper prior on $\boldsymbol{Z}$ would automatically select the optimal number of latent components. However, the model would probably lose its conjugacy, so that variational inference would no longer be practicable, and we would have to resort to sampling methods of inference.

Note that up to permutation of the latent components and sign flips, the model is identifiable thanks to the priors over $\boldsymbol{P}$ and $\boldsymbol{Q}$, even when neither $\boldsymbol{Q}$ nor the latent components are forced to be orthonormal. Furthermore, although the model is unidentifiable with respect to permutations of the latent components, due to the initialization of $\boldsymbol{Z}$ (step 1 of the algorithm in section 2), the method will always produce the same order in the latent components across different runs.

Future developments can involve a Markovian consideration of the time dynamics. By means of this extension, a connection between PLS and an input-output linear dynamical system (Beal, 2003) can be established.

Appendix A

We now formulate the variational update equations for $\boldsymbol{\Psi}^{-1}$, $\boldsymbol{Q}$, $\boldsymbol{\Gamma}^{-1}$. For a full matrix $\boldsymbol{\Psi}^{-1}$, we have a Wishart distribution,

$$F(\boldsymbol{\Psi}^{-1}) = \mathcal{W}(\boldsymbol{\Psi}^{-1}; \tilde{\boldsymbol{C}}^{-1}, \tilde{\kappa}), \tag{A.1}$$

with $\tilde{\boldsymbol{C}}^{-1} = (\boldsymbol{Y}'\boldsymbol{Y} + E[\boldsymbol{Q}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{Q}] - \boldsymbol{Y}'\boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{\mu}_{\boldsymbol{Q}} - \boldsymbol{\mu}'_{\boldsymbol{Q}}\boldsymbol{\mu}'_{\boldsymbol{Z}}\boldsymbol{Y} + \boldsymbol{C}^{-1})^{-1}$ and $\tilde{\kappa} = \kappa + N$.

For diagonal $\boldsymbol{\Psi}^{-1}$, we have the diagonal components to be gamma distributed:

$$F(\Psi_{jj}^{-1}) = \mathcal{G}(\Psi_{jj}^{-1}; \tilde{\kappa}, \tilde{C}_{jj}^{-1}), \tag{A.2}$$

with $\tilde{\kappa} = \kappa + \frac{N}{2}$ and $\tilde{C}_{jj}^{-1} = \frac{1}{2}(\boldsymbol{Y}_{\cdot j}\boldsymbol{Y}'_{\cdot j} + E[\boldsymbol{q}'_j\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{q}_j] - 2\boldsymbol{Y}'_{\cdot j}\boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{\mu}_{\boldsymbol{q}_j}) + C_{jj}^{-1}$.

For $\boldsymbol{Q}$, we have, in the general case, a kq-dimensional gaussian distribution,

$$F(\boldsymbol{Q}) = \mathcal{N}(\boldsymbol{Q}; \boldsymbol{\mu}_{\boldsymbol{Q}}, \boldsymbol{S}_{\boldsymbol{Q}}), \tag{A.3}$$

whose parameters can be reconstructed from $\boldsymbol{S}_{\tilde{\boldsymbol{q}}} = (E[\boldsymbol{\Gamma}^{-1}] \otimes \boldsymbol{I}_q + E[\boldsymbol{Z}'\boldsymbol{Z}] \otimes E[\boldsymbol{\Psi}^{-1}])^{-1}$ and $\boldsymbol{\mu}_{\tilde{\boldsymbol{q}}} = \boldsymbol{S}_{\tilde{\boldsymbol{q}}}(E[\boldsymbol{Z}'\boldsymbol{Z}] \otimes E[\boldsymbol{\Psi}^{-1}])\tilde{\boldsymbol{q}}^*$, where $\tilde{\boldsymbol{q}}$ is defined like $\tilde{\boldsymbol{p}}$ and $\tilde{\boldsymbol{q}}^*$ is the concatenation of the rows of $(\boldsymbol{\mu}'_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\boldsymbol{\mu}'_{\boldsymbol{Z}}\boldsymbol{Y}$. If we choose to factorize $F(\boldsymbol{Q})$, which follows from taking $\boldsymbol{\Psi}^{-1}$ to be diagonal, we have

$$F(\boldsymbol{q}_j) = \mathcal{N}(\boldsymbol{q}_j; \boldsymbol{\mu}_{\boldsymbol{q}_j}, \boldsymbol{S}_{\boldsymbol{q}_j}), \tag{A.4}$$

with $\boldsymbol{S}_{\boldsymbol{q}_j} = (E[\boldsymbol{\Gamma}^{-1}] + E[\Psi_{jj}^{-1}]E[\boldsymbol{Z}'\boldsymbol{Z}])^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{q}_j} = E[\Psi_{jj}^{-1}]\boldsymbol{S}_{\boldsymbol{q}_j}\boldsymbol{\mu}'_{\boldsymbol{Z}}\boldsymbol{Y}_{\cdot j}$

With regard to $\boldsymbol{\Gamma}_j^{-1}$, we have

$$F(\boldsymbol{\Gamma}^{-1}) = \mathcal{W}(\boldsymbol{\Gamma}^{-1}; \tilde{\boldsymbol{D}}^{-1}, \tilde{\varsigma}), \tag{A.5}$$

with $\tilde{\boldsymbol{D}}^{-1} = (E[\boldsymbol{Q}\boldsymbol{Q}'] + \boldsymbol{D}^{-1})^{-1}$ and $\tilde{\varsigma} = \varsigma + q$.

Appendix B

In this appendix, we derive a variational lower bound of the evidence from model 1. This can be used to monitor the inference process and check convergence. The lower bound is defined as

$$\begin{aligned}
\mathcal{L} = {} & E[\ln P(\boldsymbol{\Sigma}^{-1})] + \sum_{l=1}^{k} E[\ln P(\boldsymbol{p}_l|\boldsymbol{\Sigma}^{-1})] + E[\ln P(\boldsymbol{\Omega}^{-1})] \\
& + \sum_{n=1}^{N} E[\ln P(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{P}, \boldsymbol{\Omega}^{-1})] + E[\ln P(\boldsymbol{\Gamma}^{-1})] + \sum_{j=1}^{q} E[\ln P(\boldsymbol{q}_j|\boldsymbol{\Gamma}^{-1})] \\
& + E[\ln P(\boldsymbol{\Psi}^{-1})] + \sum_{n=1}^{N} E[\ln P(\boldsymbol{y}_n|\boldsymbol{z}_n, \boldsymbol{Q}, \boldsymbol{\Psi}^{-1})] + \sum_{l=1}^{k} E[\ln F(\boldsymbol{p}_l)]
\end{aligned}$$

$$+ E[\ln F(\boldsymbol{\Omega}^{-1}, \boldsymbol{\Sigma}^{-1})] + \sum_{n=1}^{N} E[\ln F(\boldsymbol{z}_n)] + \sum_{j=1}^{q} E[\ln F(\boldsymbol{q}_j)]$$

$$+ E[\ln F(\boldsymbol{\Psi}^{-1}, \boldsymbol{\Gamma}^{-1})]. \tag{B.1}$$

Particularizing for full matrices $\boldsymbol{\Omega}^{-1}$ and $\boldsymbol{\Psi}^{-1}$, we have

$$E[\ln P(\boldsymbol{\Sigma}^{-1})]$$

$$= -\frac{\nu}{2}\ln|\boldsymbol{B}^{-1}| - \ln\left(2^{\nu p/2}\,\pi^{p(p-1)/4}\prod_{i=1}^{p} G\Big(\frac{\nu+1-i}{2}\Big)\right)$$

$$+ \frac{\nu-p-1}{2}\left(\sum_{i=1}^{p}\psi\Big(\frac{\tilde{\nu}+1-i}{2}\Big) + p\ln 2 + \ln\big|\tilde{\boldsymbol{B}}^{-1}\big|\right) - \frac{\tilde{\nu}}{2}\,Tr(\boldsymbol{B}\tilde{\boldsymbol{B}}^{-1}),$$

$$E[\ln P(\boldsymbol{p}_l|\boldsymbol{\Sigma}^{-1})]$$

$$= -\frac{p}{2}\ln(2\pi) + \frac{1}{2}\left(\sum_{i=1}^{p}\psi\Big(\frac{\tilde{\nu}+1-i}{2}\Big) + p\ln 2 + \ln\big|\tilde{\boldsymbol{B}}^{-1}\big|\right)$$

$$- \frac{\tilde{\nu}}{2}\,\boldsymbol{\mu}'_{p_l}\tilde{\boldsymbol{B}}^{-1}\boldsymbol{\mu}_{p_l} - \frac{\nu+k}{2}\,Tr\big(\boldsymbol{S}_{p_l}\tilde{\boldsymbol{B}}^{-1}\big),$$

$$E[\ln P(\boldsymbol{\Omega}^{-1})]$$

$$= -\frac{\iota}{2}\ln|\tilde{\mathbf{A}}^{-1}| - \ln\left(2^{\iota k/2}\,\pi^{k(k-1)/4}\prod_{l=1}^{k} G\Big(\frac{\iota+1-l}{2}\Big)\right)$$

$$+ \frac{\iota-k-1}{2}\left(\sum_{l=1}^{k}\psi\Big(\frac{\tilde{\iota}+1-l}{2}\Big) + k\ln 2 + \ln\big|\tilde{\mathbf{A}}^{-1}\big|\right) - \frac{\tilde{\iota}}{2}\,Tr(\boldsymbol{A}\tilde{\mathbf{A}}^{-1}),$$

$$E[\ln P(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{P}, \boldsymbol{\Omega}^{-1})]$$

$$= -\frac{p}{2}\ln(2\pi) + \frac{1}{2}\left(\sum_{l=1}^{k}\psi\left(\frac{\tilde{\iota}+1-l}{2}\right) + k\ln 2 + \ln\big|\tilde{\mathbf{A}}^{-1}\big|\right)$$

$$- \frac{\tilde{\iota}}{2}\boldsymbol{x}'_n\left(\boldsymbol{\mu}_{\boldsymbol{P}}\tilde{\mathbf{A}}^{-1}\boldsymbol{\mu}'_{\boldsymbol{P}} + \sum_{l=1}^{k}\tilde{A}_{ll}^{-1}\boldsymbol{S}_{p_l}\right)\boldsymbol{x}_n,$$

$$E[\ln P(\boldsymbol{y}_n|\boldsymbol{z}_n, \boldsymbol{Q}, \boldsymbol{\Psi}^{-1})]$$

$$= -\frac{q}{2}\ln(2\pi) + \frac{1}{2}\left(\sum_{j=1}^{q}\psi\left(\frac{\tilde{\kappa}+1-j}{2}\right) + q\ln 2 + \ln\big|\tilde{\boldsymbol{C}}^{-1}\big|\right)$$

$$-\frac{\tilde{\kappa}}{2}\boldsymbol{\mu}'_{z_n}\left(\boldsymbol{\mu}_Q\tilde{\boldsymbol{C}}^{-1}\boldsymbol{\mu}'_Q+\sum_{j=1}^{q}\tilde{C}_{jj}^{-1}\boldsymbol{S}_{q_j}\right)\boldsymbol{\mu}_{z_n}$$

$$-\frac{\tilde{\kappa}}{2}Tr\left(\boldsymbol{S}_{z_n}\left(\boldsymbol{\mu}_Q\tilde{\boldsymbol{C}}^{-1}\boldsymbol{\mu}'_Q+\sum_{j=1}^{q}\tilde{C}_{jj}^{-1}\boldsymbol{S}_{q_j}\right)\right)-\frac{\tilde{\kappa}}{2}\boldsymbol{y}'_n\tilde{\boldsymbol{C}}^{-1}\boldsymbol{y}_n+\tilde{\kappa}\boldsymbol{y}'_n\tilde{\boldsymbol{C}}^{-1}\boldsymbol{\mu}'_Q\boldsymbol{\mu}_{z_n},$$

where $G(\cdot)$ and $\psi(\cdot)$ are the gamma and digamma functions. The expressions for $E[\ln P(\boldsymbol{\Gamma}^{-1})]$, $E[\ln P(\boldsymbol{q}_j|\boldsymbol{\Gamma}^{-1})]$ and $E[\ln P(\boldsymbol{\Psi}^{-1})]$ are analogous to $E[\ln P(\boldsymbol{\Sigma}^{-1})]$, $E[\ln P(\boldsymbol{p}_l|\boldsymbol{\Sigma}^{-1})]$, and $E[\ln P(\boldsymbol{\Omega}^{-1})]$, respectively, and are not shown.

The rest of the terms in equation B.1 correspond to the negative entropies of the $F(\cdot)$ distributions—for example:

$$E[\ln F(\boldsymbol{p}_l)]=\frac{1}{2}\ln|\boldsymbol{S}_{p_l}|+\frac{p}{2}(1+\ln(2\pi)),$$

$$E[\ln F(\boldsymbol{\Omega}^{-1},\boldsymbol{\Sigma}^{-1})]$$

$$=-\frac{\iota}{2}\ln|\tilde{\boldsymbol{\Lambda}}^{-1}|-\ln\left(2^{0.5\tilde{\iota}k}\,\pi^{k(k-1)/4}\prod_{l=1}^{k}G\left(\frac{\tilde{\iota}+1-l}{2}\right)\right)$$

$$-\frac{\tilde{\iota}-k-1}{2}\left(\sum_{l=1}^{k}\psi\left(\frac{\tilde{\iota}+1-l}{2}\right)+k\ln 2+\ln\left|\tilde{\boldsymbol{\Lambda}}^{-1}\right|\right)+\frac{\tilde{\iota}k}{2}$$

$$-\frac{\nu}{2}\ln|\tilde{\boldsymbol{B}}^{-1}|-\ln\left(2^{0.5\tilde{\nu}p}\,\pi^{p(p-1)/4}\prod_{i=1}^{p}G\left(\frac{\tilde{\nu}+1-i}{2}\right)\right)$$

$$-\frac{\tilde{\nu}-p-1}{2}\left(\sum_{i=1}^{p}\psi\left(\frac{\tilde{\nu}+1-i}{2}\right)+p\ln 2+\ln\left|\tilde{\boldsymbol{B}}^{-1}\right|\right)+\frac{\tilde{\nu}p}{2},$$

$$E[\ln F(\boldsymbol{z}_n)]=\frac{1}{2}\ln|\boldsymbol{S}_{z_n}|+\frac{k}{2}(1+\ln(2\pi)).$$

The variational lower bound for other variations of the method can be easily computed following the same line of argument.

Acknowledgments

This work has been partially supported by projects TIN2010-20900-C04-04 and Cajal Blue Brain of the Spanish Ministry of Science and Innovation (MINECO).

References

Ando, R., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, *6*, 1817–1853.

Argyriou, A., Evgeniou, T., & Pontil, M. (2006). Multi-task feature learning. In B. Schölkopf, B. Platt, & T. Hoffmann (Eds.), *Advances in neural information processing systems, 19* (pp. 41–48). Cambridge, MA: MIT Press.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*, 166–173.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. Unpublished doctoral dissertation, University College London.

Bishop, C. (1998). Bayesian principal components. In M. S. Keams, S. Sola, & D. A Cohn (Eds.), *Advances in neural information processing systems, 11*. Cambridge, MA: MIT Press.

Brown, P. J., & Zidek, J. V. (1980). Adaptive multivariate ridge regression. *Annals of Statistics*, *8*, 64–74.

Chun, H., & Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B*, *72*, 3–25.

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*, 251–263.

Frank, I. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*, 109–135.

Fujiwara, Y., Miyawaki, Y., & Kamitani, Y. (2009). Estimating image bases for visual image reconstruction from human brain activity. In Y. Bengio, P. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems*, *22*. Cambridge, MA: MIT Press.

Ghahramani, Z., & Beal, M. J. (2000). Variational inference for Bayesian mixtures of factor analysers. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems*, *12*. Cambridge, MA: MIT Press.

Goutis, C. (1996). Partial least squares yields shrinkage estimators. *Annals of Statistics*, *24*, 816–824.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, *16*, 2639–2664.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: Data mining, inference, and predictions* (2nd ed.). New York: Springer.

Izenman, A. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, *5*, 248–264.

Jaakkola, T. S. (2001). Tutorial on variational approximation methods. In M. Opper & D. Saad (Eds.), *Advanced mean field methods: Theory and practice*. Cambridge, MA: MIT Press.

Lindgren, F., Geladi, P., & Wold, S. (1993). The kernel algorithm for PLS. *Journal of Chemometrics*, *7*, 45–59.

Nakajima, S., Sugiyama, M., & Babacan, D. (2011). On Bayesian PCA: Automatic dimensionality selection and analytic solution. In *37th International Conference on Machine Learning*. International Machine Learning Society.

Obozinski, G., Wainwright, M. J., & Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, *39*, 1–47.

Rosipal, R., & Krämer, N. (2006). Overview and recent advances in partial least squares. *Lecture Notes in Computer Science, 3940*, 34–51.

Rosipal, R., & Trejo, L. J. (2001). Kernel partial least squares regression in reproducing kernel Hilbert space. *Journal of Machine Learning Research*, *2*, 97–123.

Rothman, A. J., Levina, E., & Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, *19*, 947–962.

van Gerven, M. A. J., Chao, Z. C., & Heskes, T. (2012). On the decoding of intracranial data using sparse orthonormalized partial least squares. *Journal of Neural Engineering*, *9*, 026017.

Virtanen, S., Klami, A., & Kaski, S. (2011). Bayesian CCA via group sparsity. In *37th International Conference on Machine Learning*. International Machine Learning Society.

Wang, C. (2007). Variational Bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, *18*, 905–910.

Wangen, L. E., & Kowalsky, B. R. (1989). A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics*, *3*, 3–20.

Wold, H. (1975). Path models with latent variables: The NIPALS approach. In H. M. Blalock (Ed.), *Quantitative sociology: International perspectives on mathematical and statistical model building* (pp. 307–357). New York: Academic Press.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*, 109–130.

Yuan, M., Ekici, A., Lu, Z., & Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B*, *69*, 329–346.

Received October 18, 2012; accepted July 13, 2013.