# Supplementary Material: Bayesian Factor Analysis for Inference on Interactions

Federico Ferrari
Department of Statistical Science, Duke University
and
David B. Dunson
Department of Statistical Science, Duke University

March 15, 2020

# 1 Grouping of Parameters with a Block Sparse $\Lambda$

Let us consider the scenario when the true number of factors is equal to 2, and the variables in $X$ can be divided in two groups. This structure is reasonable when we have measurements of different chemicals that are breakdown products of the same exposure, for example PCB metabolites (Longnecker et al., 2001). Then the a priori covariance between main effects is equal to $Cov(\beta_g, \beta_h | g, h \in S_r) = \frac{\lambda_h \lambda_g}{(\sum_{j \in S_r}^{p} \lambda_j^2 + \sigma^2)^2}$, for $r = 1, 2$, when the chemicals belong to the same group, and is zero otherwise. Hence, in this case, $A^T A$ is block diagonal. On the other hand, assume that the number of factors is equal to the number of covariates, with $\Lambda$ being diagonal. In this case the induced covariance on $\beta$ is diagonal with elements $\frac{\lambda_j^2}{(\lambda_j^2+\sigma^2)^2}$. In general when there are $l$ groups, the variance of $\beta_h$, with $h \in S_r$, is equal to $\frac{\lambda_h^2}{(\sum_{j \in S_r} \lambda_j^2+\sigma^2)^2}$. Hence, the variance of $\beta$ is lower with respect to the independent case since we are borrowing strength and information from the other covariates within the same group.

Let us now focus on the symmetric matrix $\Omega$ of dimension $k$, letting $\nu(\Omega)$ be the vector of lower triangular elements of $\Omega$. Define the duplication matrix $D_k$ as the $k^2 \times \frac{k(k+1)}{2}$ matrix such that $D_k \nu(\Omega) = vec(\Omega)$, see Magnus (1988) as a reference. The duplication matrix can be easily calculated for orders 2 and 3, whereas the R package `matrixcalc` provides the duplication matrix for higher orders. We are interested in the distribution of $\Omega_X = A^T \Omega A$. Notice that

$$vec(\Omega_X) = vec(A^T \Omega A) = (A^T \otimes A^T) vec(\Omega) = (A^T \otimes A^T) D_k \nu(\Omega).$$

We choose a normal prior on pairwise interactions, i.e., $\nu(\Omega) \sim N_{\frac{k(k+1)}{2}}(0, I_{\frac{k(k+1)}{2}})$, so that $vec(\Omega_X) \sim N_{p^2}(0, (A^T \otimes A^T) D_k D_k^T (A \otimes A))$. Computing the covariance $(A^T \otimes A^T) D_k D_k^T (A \otimes A)$ of the induced Normal prior on the matrix containing the pairwise inter-

actions $vec(\Omega_X) = (\omega_{1,1}, \omega_{1,2}, \cdots, \omega_{4,3}, \omega_4)$, we find that the variables are divided in three groups: $(\omega_{1,1}, \omega_{1,2}, \omega_{2,2})$, $(\omega_{3,3}, \omega_{3,4}, \omega_{4,4})$ and $(\omega_{1,3}, \omega_{1,4}, \omega_{2,3}, \omega_{2,4})$. The quadratic effect of the first two covariates are correlated with each other and with the interaction between them. The same holds for the variables loading on the second factor. Finally, the third group contains the interactions between one variable loading on the first factor and one variable loading on the second factor. In general, with $p$ variables and $k$ factors, we will have in total $k + \binom{k}{2}$ groups. In particular, the first $k$ groups will be made by the interactions between variables loading on the same factor. On the other hand, we have groups for the interactions between variables loading on different factors, as in the previous example with $S_1$ and $S_2$.

## 2 Complexity Gains

Inference under existing approaches for Bayesian linear modeling for pairwise interactions when $p$ is moderately high is typically computationally infeasible. In fact the complexity per iteration of Gibbs sampling is $\mathcal{O}(np^4 + p^6)$ and the storage is of order $\mathcal{O}(p^2)$. This is without considering any heredity structure. On the other hand, with model (1) we just need samples of $\Psi$, $\Lambda$, $\omega$ and $\Omega$ to compute main effects and interactions of $X$ on $y$ thanks to *Proposition 1.* The complexity per iteration of Gibbs sampling is $\mathcal{O}(k^3p + npk)$, where $k$ is the number of factors. In our motivating applications, we have $n > p > k$. Further, the storage complexity is only $\mathcal{O}(pk)$ since we only need to save the samples of $\Lambda, \Phi, \omega$ and $\Omega$.

The computational complexity could be further reduced using the algorithm of Sabnis et al. (2016), which allows one to distribute the covariance matrix estimation to multiple cores, efficiently using a divide and conquer strategy. Let $g \geq 1$ denote the number of cores

at our disposal and assume that $p$ is a multiple of $g$. Letting $p_g = \frac{p}{g}$, the computational complexity becomes $\mathcal{O}(k^3 p_g + n p_g k)$. If we want to estimate the interactions up to the $Q^{th}$ order, the computational complexity becomes $\mathcal{O}(k^3(p+Q) + npk)$. Moreover, the storage complexity is $\mathcal{O}(p(k+Q))$.

# 3 Proofs

## 3.1 Proof of Proposition 2

As $n$ grows, we have that the posterior distribution of $(\Lambda, \Psi)$ concentrates on the model that is closest to the true data-generating model in Kullback-Leibler divergence, see (Berk et al., 1966).

$$KL((\Lambda_0, \Psi_0); (\Lambda, \Psi)) = tr\left((\Lambda\Lambda^T + \Psi)^{-1}(\Lambda_0\Lambda_0^T + \Psi_0)\right) - p + log\left(\frac{|\Lambda\Lambda^T + \Psi|}{|\Lambda_0\Lambda_0^T + \Psi_0|}\right)$$

where $KL((\Lambda_0, \Psi_0), (\Lambda, \Psi))$ denotes the Kullback-Leibler divergence between $p(X, y|\Lambda_0, \Psi_0)$ and $p(X, y|\Lambda, \Psi_0)$. Let $(\Lambda^*, \Psi^*) = arg\ inf\Big[d(\Lambda, \Psi)\Big]$. Now, $\Lambda_0\Lambda_0^T$ is symmetric and positive definite, so it admits an EigenDecomposition, i.e.:

$$\Lambda_0\Lambda_0^T + s_0 I_p = Q_0\Sigma_0 Q_0^{-1} + s_0 I_p = Q_0(\Sigma_0 + s_0 I_p)Q_0^{-1},$$

where $Q_0$ is the $p \times p$ matrix containing the eigenvectors of $\Lambda_0\Lambda_0^T$ and $\Sigma_0$ is a diagonal matrix containing the eigeinvalues, i.e. $diag(\Sigma_0) = (v_1, \cdots, v_{k_0}, 0 \cdots, 0)$. Define $\Lambda_1\Lambda_1^T$ as the best $k^{th}$ rank approximation to $\Lambda_0\Lambda_0^T$, when the approximation is based on the Frobenious norm. From the Eckart-Young theorem, we know that $\Lambda_1\Lambda_1^T = Q_0\Sigma_1 Q_0^{-1}$ where $diag(\Sigma_1) = (v_1, \cdots, v_k, 0 \cdots, 0)$. By definition of $(\Lambda^*, \Psi^*)$:

$KL((\Lambda_0, \Psi_0); (\Lambda^*, \Psi^*)) \leq KL((\Lambda_0, \Psi_0); (\Lambda_1, \Psi_0)) =$

$$= tr\Bigg((Q_0(\Sigma_1 + s_0 I_p)Q_0^{-1})^{-1}(Q_0(\Sigma_0 + s_0 I_p)Q_0^{-1})\Bigg) - p + log\Bigg(\frac{|Q_0(\Sigma_1 + s_0 I_p)Q_0^{-1}|}{|Q_0(\Sigma_0 + s_0 I_p)Q_0^{-1}|}\Bigg) =$$

$$= tr\Bigg((\Sigma_1 + s_0 I_p)^{-1}((\Sigma_0 + s_0 I_p))\Bigg) - p + \sum_{j=k+1}^{k_0}\Bigg(log(s_0) - log(v_j + s_0)\Bigg) \leq$$

$$\leq k + \sum_{j=k+1}^{k_0}\frac{s_0}{s_0 + v_j} + (p - k_0) - p =$$

$$= \sum_{j=k+1}^{k_0}\Bigg(\frac{s_0 + v_j}{s_0} - 1\Bigg) = \sum_{j=k+1}^{k_0}\frac{v_j}{s_0}$$

## 3.2 Proof of Proposition 3

Let $p_0(X, y) = p(X, y|\Theta_0) = \int p(X, y|\Theta_0, \eta)p(\eta)d\eta$ where $\Theta_0 = (\omega_0, \Omega_0, \sigma_0^2, \Phi_0)$ and let $p'(X, y) = p(X, y|\eta')$ for a given vector $\eta'$. Also, we have that $p_0(X, y) = C_0 k_0(X, y)$, where $k_0(X, y)$ is the kernel of a Multivariate normal distribution with parameters $\Theta_0$. We are interested in computing:

$$KL(p_0; p') = \int p_0(X, y)\log(\frac{p_0(X, y)}{p'(X, y)})dXdy$$

Let us focus on the $p_0(X, y)$:

$$p_0(X, y) = \int p_0(X, y|\eta)p(\eta)d\eta =$$

$$= \int_{B_{1-\epsilon}} p_0(X, y|\eta)p(\eta)d\eta + \int_{B_{1-\epsilon}^C} p_0(X, y|\eta)p(\eta)d\eta$$

Where $B_{1-\epsilon}$ is a closed ball such that $p(\eta \in B_{1-\epsilon}) = 1 - \epsilon$ according to the prior $p(\eta)$. Now, on the closed ball $B_{1-\epsilon}$ the function $p_0(X, y|\eta)$ has a supremum which we denote $\eta^* = \eta(y, X, \Theta_0) = argsup_{\eta \in B_{1-\epsilon}} p_0(X, y|\eta)$. Also recall that $p_0(X, y|\eta) = C_0 k_0(X, y|\eta)$ where $k_0(X, y|\eta) \leq 1$.

$$p_0(X, y) = \int_{B_{1-\epsilon}} p_0(X, y|\eta)p(\eta)d\eta + \int_{B_{1-\epsilon}^C} p_0(X, y|\eta)p(\eta)d\eta \leq$$

$$\leq \int_{B_{1-\epsilon}} p_0(X, y|\eta^*)p(\eta)d\eta + \int_{B^C_{1-\epsilon}} C_0 p(\eta)d\eta =$$
$$= p_0(X, y|\eta^*)(1-\epsilon) + C_0\epsilon$$

We now need to take the logarithm of the expression above, recall *log sum inequality*:

$$\log(\frac{a_1+a_2}{b_1+b_2}) \leq \frac{a_1}{a_1+a_2}\log(\frac{a_1}{b_1}) + \frac{a_2}{a_1+a_2}\log(\frac{a_2}{b_2}) \leq$$
$$\leq \log(\frac{a_1}{b_1}) + \log(\frac{a_2}{b_2})$$

We apply the *log sum inequality* with $a_1 = p_0(X, y|\eta^*)(1-\epsilon)$, $a_2 = C_0\epsilon$, $b_1 = 1-\epsilon^*$ and $b_2 = \epsilon^*$, so we have that:

$$\log(\frac{p_0(X, y|\eta^*)(1-\epsilon) + C_0\epsilon}{1-\epsilon^*+\epsilon^*}) \leq \log(p_0(X, y|\eta^*)\frac{1-\epsilon}{1-\epsilon^*}) + \log(\frac{C_0\epsilon}{\epsilon^*}) =$$
$$= \log(p_0(X, y|\eta^*)) + \log(\frac{1-\epsilon}{1-\epsilon^*}) + \ \log(\frac{C_0\epsilon}{\epsilon^*}) \leq$$
$$\leq \log(p_0(X, y|\eta^*)) - \log(1-\epsilon^*) + \log(\frac{C_0\epsilon}{\epsilon^*})$$

We can choose $\epsilon^*$ s.t. $-log(1-\epsilon^*) \leq \epsilon_1$ and $\epsilon = \frac{\epsilon^*}{C_0}$ so that the last term in the above expression is equal to zero. We can also choose $\epsilon \leq \frac{\epsilon^*}{C_0}$ and we would have $log(\frac{C_0\epsilon}{\epsilon^*}) \leq 0$. Finally we have that:

$$KL(p_0; p') \leq \int p_0(X, y)\log(\frac{p_0(X, y|\eta^*)}{p(X, y|\eta')})dXdy + \epsilon_1$$

We can now compute the above integral, find it below multiplied by 2:

$$\log\left(\frac{\prod_{j=1}^p \psi_j^0}{\prod_{j=1}^p \psi_j}\right) + \left[(\Lambda\eta')^T\Psi^{-1}(\Lambda\eta') - (\Lambda_0\eta^*)^T\Psi_0^{-1}(\Lambda_0\eta^*)\right] + tr\left((\Psi^{-1} - \Psi_0^{-1})\text{Cov}_0(X)\right) +$$
$$+ \log\left(\frac{\sigma_0^2}{\sigma_0}\right) + \mathbb{E}_0(y^2)\left(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2}\right) + 2\mathbb{E}_0(y)\left[\frac{\eta^*\omega_0 + \eta^*\Omega_0\eta^*}{\sigma_0^2} - \frac{\eta'\omega + \eta'\Omega\eta'}{\sigma^2}\right] +$$

$$+\left[\frac{(\eta'\omega+\eta'\Omega\eta')^2}{\sigma^2}-\frac{(\eta^*\omega_0+\eta^*\Omega_0\eta^*)^2}{\sigma_0^2}\right]$$

We have that $2\mathbb{E}_0(X^T)\left[\Psi_0^{-1}\Lambda_0\eta^*-\Psi^{-1}\Lambda\eta'\right]=0$. These are all continuous functions of $\Theta,\Theta,\eta^*$ and $\eta'$, so we can choose $\Theta$ and $\eta'$ such that the above expression is $\leq\epsilon_1$ so that $KL(p_0;p')\leq 2\epsilon_1$. In particular there exist $\delta$ such that this holds for any $\eta'\in D_\delta=\{\eta:||\eta-\eta^*||_2\leq\delta\}$. Hence we have that $\Phi(D_\delta)>0$, where $\Phi$ is the multivariate normal distribution and we can apply *Proposition 6.28* of Ghosal and Van der Vaart (2017) to get the result.

# References


Berk, R. H. et al. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics 37*(1), 51–58.

Ghosal, S. and A. Van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*, Volume 44. Cambridge University Press.

Longnecker, M. P., M. A. Klebanoff, H. Zhou, and J. W. Brock (2001). Association between maternal serum concentration of the ddt metabolite dde and preterm and small-for-gestational-age babies at birth. *The Lancet 358*(9276), 110–114.

Magnus, J. R. (1988). *Linear structures*, Volume 42. Griffin.

Sabnis, G., D. Pati, B. Engelhardt, and N. Pillai (2016). A divide and conquer strategy for high dimensional Bayesian factor models. *ArXiv Statistics Methodology*.

# Figures and Tables

| | | HierNet | FAMILY | PIE | RAMP | FIN |
|---|---|---|---|---|---|---|
| factor | test err | 1 | 5.011 | 4.080 | 29.464 | 1.207 |
| | FR | 1.317 | 1 | 1.369 | 1.098 | 1.021 |
| | main MSE | 1.557 | 1 | 1.385 | 3.383 | 1.255 |
| | TP main | 0.974 | 1 | 0.053 | 0.056 | 0.741 |
| | TN main | 0.027 | 0.002 | 0.975 | 0.961 | 0.467 |
| | TP int | 0.135 | 0.858 | 0.072 | 0.002 | 0.662 |
| | TN int | 0.911 | 0.171 | 0.951 | 0.998 | 0.388 |
| linear | test err | 1 | 1.547 | 2.100 | 4.717 | 1.781 |
| | FR | 1.167 | 1 | 1.165 | 1.102 | 1.045 |
| | main MSE | 14.006 | 1 | 1.873 | 3.267 | 1.827 |
| | TP main | 1 | 1 | 0.071 | 0.034 | 0.114 |
| | TN main | 0 | 0 | 0.945 | 0.968 | 0.913 |
| | TP int | 0.332 | 0.798 | 0.068 | 0.002 | 0.147 |
| | TN int | 0.866 | 0.280 | 0.966 | 0.998 | 0.902 |
| independent | test err | 1 | 1.072 | 1.251 | 1.322 | 1.239 |
| | FR | 1.211 | 1 | 1.112 | 1.140 | 1.105 |
| | main MSE | 9.340 | 1.565 | 1.295 | 1.378 | 1 |
| | TP main | 1 | 1 | 0.091 | 0.013 | 0 |
| | TN main | 0 | 0 | 0.947 | 0.996 | 1 |
| | TP int | 0.516 | 0.861 | 0.082 | 0 | 0.0002 |
| | TN int | 0.810 | 0.308 | 0.991 | 1 | 1 |

Table 1: Results from simulation study with $p = 50$ and *dense* $\Omega_0$ in the three scenarios: factor, linear and independent for $n = 500$. We computed test error, Frobenious norm, MSE for main effects, percentage of true positives and true negatives for main effects and interactions for Hiernet, Family, PIE, RAMP and FIN model with $a = 0.5$ across 50 simulations. Test error, FR, and main MSE are presented as ratios compared to the best performing model.

| | | HierNet | FAMILY | PIE | RAMP | FIN |
|---|---|---|---|---|---|---|
| factor | test error | 2.755 | 16.096 | 3.381 | 24.331 | 1 |
| | FR | 1 | 1.184 | 1.103 | 1.358 | 1.168 |
| | main MSE | 1.367 | 1.209 | 1.183 | 1.853 | 1 |
| | TP main | 0.825 | 0.958 | 0.415 | 0.264 | 0.812 |
| | TN main | 0.225 | 0.067 | 0.890 | 0.934 | 0.606 |
| | TP int | 0.631 | 0.940 | 0.485 | 0.014 | 0.645 |
| | TN int | 0.966 | 0.264 | 0.967 | 0.995 | 0.893 |
| linear | test error | 1 | 5.894 | 1.168 | 8.743 | 3.281 |
| | FR | 1.122 | 1.990 | 1 | 2.154 | 2.114 |
| | main MSE | 1 | 1.751 | 1.614 | 2.148 | 1.674 |
| | TP main | 1 | 0.900 | 0.496 | 0.342 | 0.696 |
| | TN main | 0.208 | 0.141 | 0.856 | 0.954 | 0.787 |
| | TP int | 0.952 | 0.875 | 0.731 | 0.014 | 0.726 |
| | TN int | 0.972 | 0.560 | 0.972 | 0.997 | 0.877 |
| independent | test error | 1 | 7.881 | 1.738 | 10.984 | 10.288 |
| | FR | 1.595 | 2.588 | 1 | 2.884 | 2.762 |
| | main MSE | 1 | 2.687 | 2.859 | 3.573 | 4.116 |
| | TP main | 1 | 1 | 0.350 | 0.062 | 0.112 |
| | TN main | 0.054 | 0.012 | 0.919 | 1 | 0.979 |
| | TP int | 1 | 0.985 | 0.753 | 0 | 0.025 |
| | TN int | 0.979 | 0.355 | 0.979 | 1 | 1 |

Table 2: Results from simulation study with $p = 50$ and *sparse* $\Omega_0$ in the three scenarios: factor, linear and independent for $n = 500$. We computed test error, Frobenious norm, MSE for main effects, percentage of true positives and true negatives for main effects and interactions for Hiernet, Family, PIE, RAMP and FIN model with $a = 0.5$ across 50 simulations. Test error, FR, and main MSE are presented as ratios compared to the best performing model.

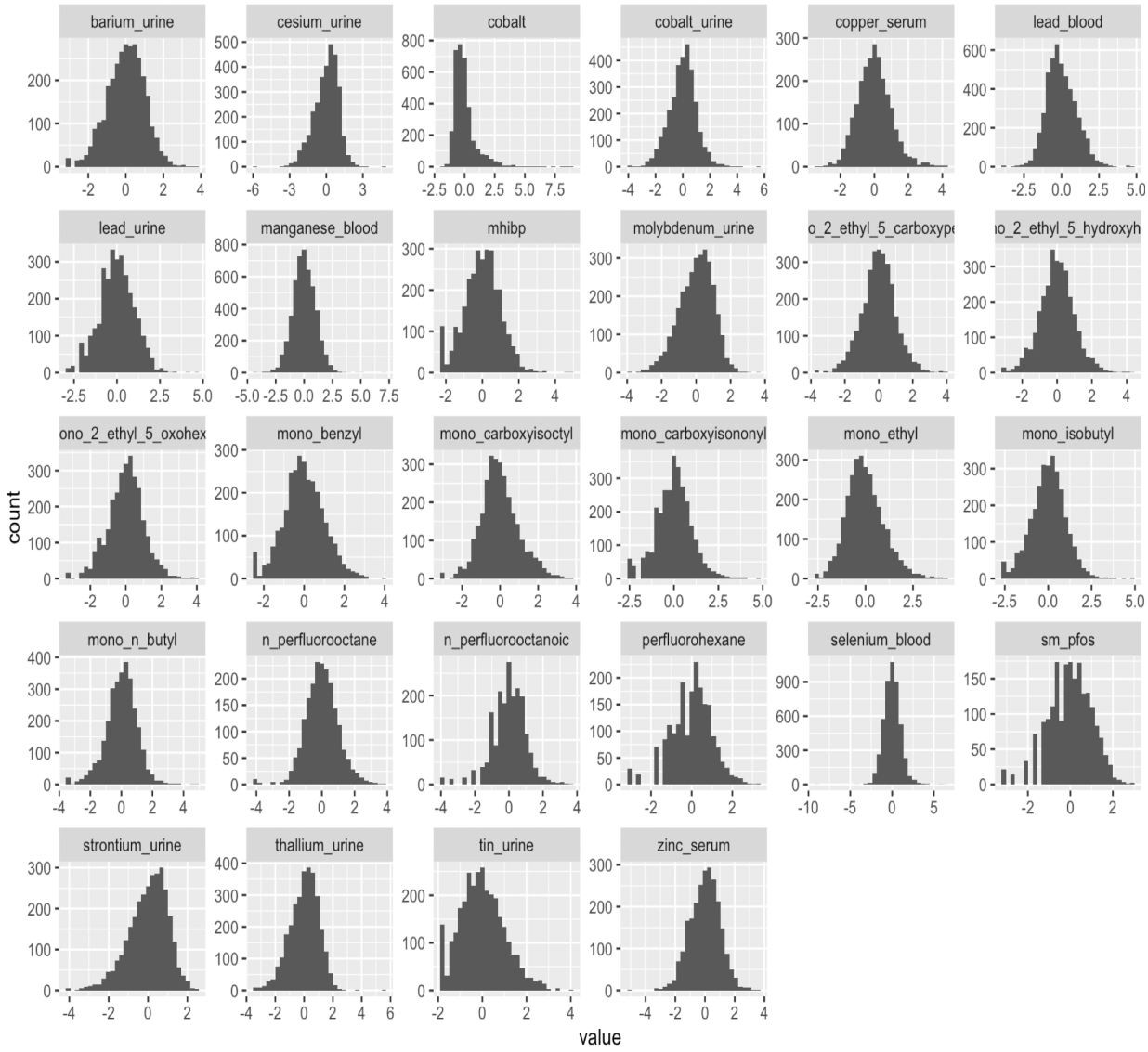


Figure 1: Histograms of the chemicals measurements included in the matrix $X$ in *Section 5*.

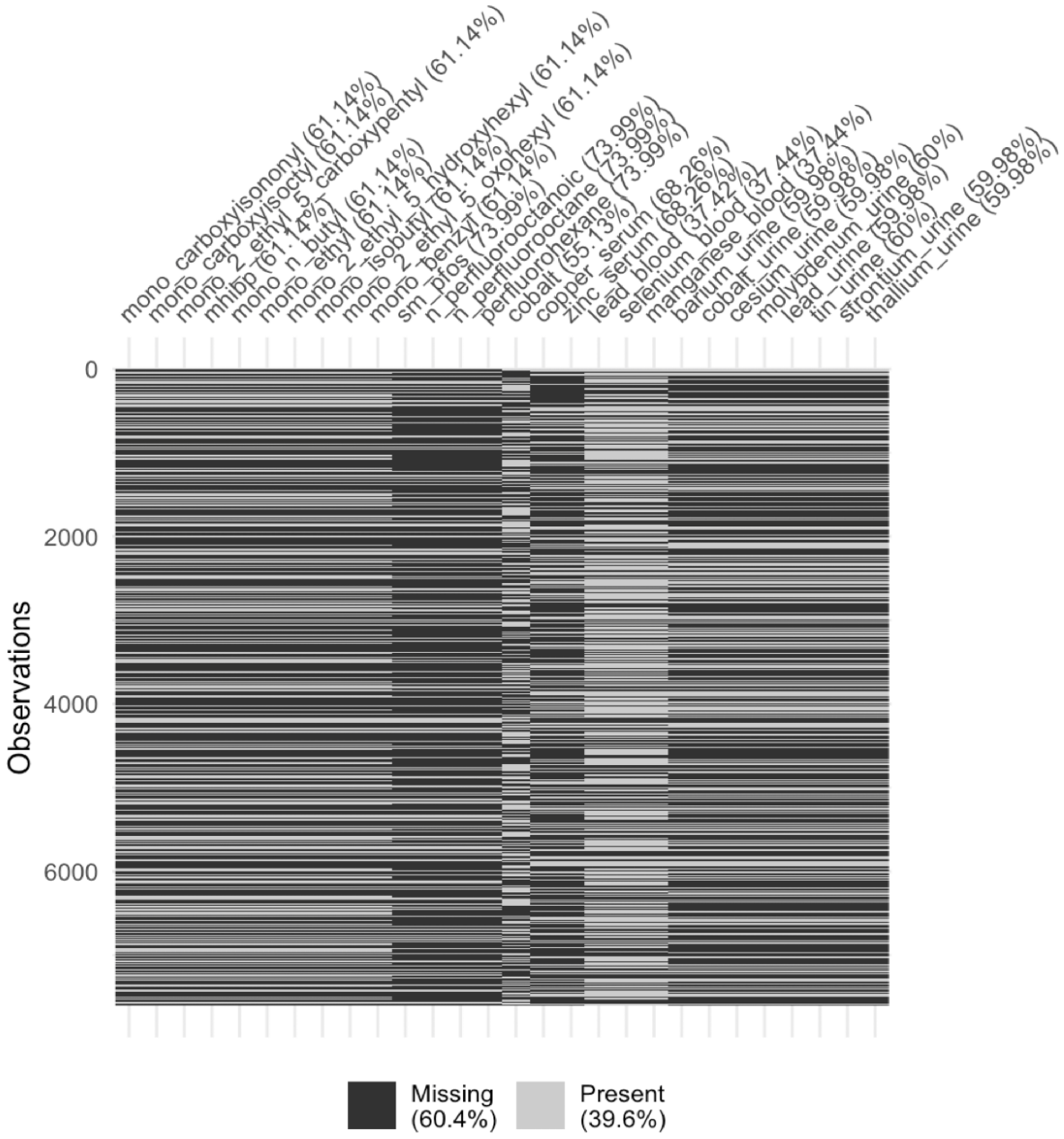


Figure 2: Pattern of Missing data in the matrix $X$ including the chemical measurements.