5

A Bayesian framework for differential proteomics analysis

5.1 Background: Bayesian inference for Gaussian-inverse-gamma conjugated priors 109
5.2 General Bayesian framework for evaluating mean differences 114
5.3 The uncorrelated case: no more multiple testing nor imputation 120
5.4 Experiments 122
5.4.1 Univariate Bayesian inference for differential analysis 123
5.4.2 The benefit of intra-protein correlation 124
5.4.3 The mirage of imputed data 126
5.4.4 Acknowledging the effect size 127
5.4.5 About protein inference 128
5.5 Conclusion and perspectives 131

In the state-of-the-art approach of Smyth (2004), as well as in our methodology described in Chapter 3, a hierarchical model is used to deduce the posterior distribution of the variance estimator for each analyte. The expectation of this distribution is then used as a moderated estimation of variance and is injected directly in the expression of the t-statistic. However, instead of relying simply on the moderated estimates, it could make sense to take advantage from a fully Bayesian approach.

The topic of missing data has been under investigation for a long time in the Bayesian community, in particular in simple cases involving conjugate priors (Dominici et al., 2000). Despite such theoretical advances, practitionners in proteomics often still rely on old fashioned tools, like t-tests, for conducting most of the differential analyses. Recently, some authors provided convenient approaches and associated implementations (Kruschke, 2013) for handling differential analysis problems with Bayesian inference. For instance, the `R` package `BEST` (standing for Bayesian Estimation Supersedes T-test) has widely contributed to the diffusion of those practices. The present chapter follows a similar idea, by taking advantage of standard results from Bayesian inference with conjugate priors in hierarchical models, to derive a methodology that is tailored to handle our multiple imputation context. Furthermore, we also aim at tackling the more general problem of multivariate differential analysis, to account for possible correlations between analytes.

By defining a hierarchical model with prior distributions both on mean and variance parameters, we aim at providing an adequate quantification of the uncertainty for differential analysis. Inference is thus performed by computing the posterior distribution for the difference of mean peptide intensity between two experimental conditions. In contrast to more flexible models that can be achieved with hierarchical structures, our choice of conjugate priors maintains analytical expressions for direcly sampling from posterior distributions without needing MCMC methods, resulting in a fast inference procedure in practice.

Section 5.1 presents well-known results about Bayesian inference for Gaussian-inverse-gamma conjugated priors. Following analogous results for the multivariate case, Section 5.2 introduces a general Bayesian framework for evaluating mean differences in our differential proteomics context. Section 5.3 provides insights on the particular case where the considered analytes are uncorrelated. Finally, Section 5.4 illustrates hands-on examples on a real proteomics dataset and highlights the benefits of such a multivariate Bayesian framework for practitioners.

5.1 Background: Bayesian inference for Gaussian-inverse-gamma conjugated priors

Before deriving our complete workflow, let us first recall some classical results from Bayesian inference that will further serve our aim. The purpose of this section is twofold. By first fully detailing proofs of results in the univariate case that are often admitted, we pave the way to the development of our subsequent contribution in a multivariate framework.

Let us assume a generative model such as:

$$y = \mu + \varepsilon,$$

where:

- $\mu \mid \sigma^2 \sim \mathcal{N}\left(\mu_0, \frac{1}{\lambda_0}\sigma^2\right)$ is the prior distribution over the mean,
- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is the error term,
- $\sigma^2 \sim \Gamma^{-1}(\alpha_0, \beta_0)$ is the prior distribution over the variance,

with $\{\mu_0, \lambda_0, \alpha_0, \beta_0\}$ an arbitrary set of prior hyper-parameters. We provide in Figure 5.1 an illustration of the hypotheses taken over such hierarchical generative model. From the

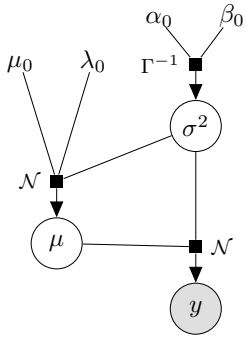


Figure 5.1: Graphical model of the hierarchical structure when assuming a Gaussian-inverse-gamma prior, conjugated with a Gaussian likelihood with unknown mean and variance.

previous hypotheses, we can deduce the likelihood of the model for a sample of observations $\boldsymbol{y} = \{y_1, \ldots, y_N\}$:

$$\begin{aligned} p(\boldsymbol{y} \mid \mu, \sigma^2) &= \prod_{n=1}^{N} p(y_n \mid \mu, \sigma^2) \\ &= \prod_{n=1}^{N} \mathcal{N}\left(y_n;\ \mu, \sigma^2\right), \end{aligned}$$

Let us recall that such assumptions consists in defining a prior Gaussian-inverse-gamma distribution, which is conjugated with the Gaussian distribution with unknown mean μ and variance σ^2. The probability density function (PDF) of such a prior distribution can be written as:

$$p(\mu, \sigma^2 \mid \mu_0, \lambda_0, \alpha_0, \beta_0) = \frac{\sqrt{\lambda_0}}{\sqrt{2\pi}} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\frac{1}{\sigma^2}\right)^{\alpha_0 + \frac{3}{2}} \exp\left(-\frac{2\beta_0 + \lambda_0(\mu - \mu_0)^2}{2\sigma^2}\right).$$

In this particular case, it is a well-known result that the inference is tractable and the posterior distribution remains a Gaussian-inverse-gamma (Murphy, 2007). Let us recall below the complete development of this derivation by identification of the analytical form

(we ignore conditioning over the hyper-parameters for convenience):

$$
\begin{aligned}
p(\mu, \sigma^2 \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \mu, \sigma^2) \times p(\mu, \sigma^2) \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mu)^2\right) \\
&\quad \times \frac{\sqrt{\lambda_0}}{\sqrt{2\pi}}\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\left(\frac{1}{\sigma^2}\right)^{\alpha_0+\frac{3}{2}} \exp\left(-\frac{2\beta_0 + \lambda_0(\mu-\mu_0)^2}{2\sigma^2}\right) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{\alpha_0+\frac{N+3}{2}} \exp\left(\underbrace{-\frac{2\beta_0 + \lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^{N}(y_n-\mu)^2}{2\sigma^2}}_{\mathcal{A}}\right).
\end{aligned}
$$

Let us introduce Lemma 5.1 below to decompose the term $\mathcal{A}$ as desired:

Lemma 5.1. *Assume a set $\boldsymbol{x_1}, \ldots, \boldsymbol{x_N} \in \mathbb{R}^q$, and note $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x_n}$ the associated average vector. For any $\boldsymbol{\mu} \in \mathbb{R}^q$:*

$$
\sum_{n=1}^{N}(\boldsymbol{x_n} - \boldsymbol{\mu})(\boldsymbol{x_n} - \boldsymbol{\mu})^\intercal = N(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\intercal + \sum_{n=1}^{N}(\boldsymbol{x_n} - \bar{\boldsymbol{x}})(\boldsymbol{x_n} - \bar{\boldsymbol{x}})^\intercal.
$$

Proof.

$$
\begin{aligned}
\sum_{n=1}^{N}(\boldsymbol{x_n} - \boldsymbol{\mu})(\boldsymbol{x_n} - \boldsymbol{\mu})^\intercal &= \sum_{n=1}^{N}\boldsymbol{x_n}\boldsymbol{x_n}^\intercal + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal - 2\boldsymbol{x_n}\boldsymbol{\mu}^\intercal \\
&= N\boldsymbol{\mu}\boldsymbol{\mu}^\intercal - 2N\bar{\boldsymbol{x}}\boldsymbol{\mu}^\intercal + \sum_{n=1}^{N}\boldsymbol{x_n}\boldsymbol{x_n}^\intercal \\
&= N\boldsymbol{\mu}\boldsymbol{\mu}^\intercal + N\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal + N\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal - 2N\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal - 2N\bar{\boldsymbol{x}}\boldsymbol{\mu}^\intercal + \sum_{n=1}^{N}\boldsymbol{x_n}\boldsymbol{x_n}^\intercal \\
&= N\left(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal - 2\bar{\boldsymbol{x}}\boldsymbol{\mu}^\intercal\right) + \sum_{n=1}^{N}\boldsymbol{x_n}\boldsymbol{x_n}^\intercal + \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal - 2\boldsymbol{x_n}\bar{\boldsymbol{x}}^\intercal \\
&= N\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right)\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right)^\intercal + \sum_{n=1}^{N}(\boldsymbol{x_n} - \bar{\boldsymbol{x}})(\boldsymbol{x_n} - \bar{\boldsymbol{x}})^\intercal.
\end{aligned}
$$

□

Applying this result in our context for $q = 1$, we obtain:

$$
\mathcal{A} = -\frac{1}{2\sigma^2}\left(2\beta_0 + \lambda_0(\mu - \mu_0)^2 + N(\bar{y} - \mu)^2 + \sum_{n=1}^{N}(y_n - \bar{y})^2\right)
$$

$$
\begin{aligned}
&= -\frac{1}{2\sigma^2}\left(2\beta_0 + \sum_{n=1}^{N}(y_n-\bar{y})^2 + (\lambda_0+N)\mu^2 - 2\mu(N\bar{y}+\lambda_0\mu_0) + N\bar{y}^2 + \lambda_0\mu_0^2\right)\\
&= -\frac{1}{2\sigma^2}\left(2\beta_0 + \sum_{n=1}^{N}(y_n-\bar{y})^2 + N\bar{y}^2 + \lambda_0\mu_0^2\right.\\
&\quad \left. + (\lambda_0+N)\left[\mu^2 - 2\mu\frac{N\bar{y}+\lambda_0\mu_0}{\lambda_0+N} + \left(\frac{N\bar{y}+\lambda_0\mu_0}{\lambda_0+N}\right)^2 - \left(\frac{N\bar{y}+\lambda_0\mu_0}{\lambda_0+N}\right)^2\right]\right)\\
&= -\frac{1}{2\sigma^2}\left(2\beta_0 + \sum_{n=1}^{N}(y_n-\bar{y})^2 + N\bar{y}^2 + \lambda_0\mu_0^2 - \frac{(N\bar{y}+\lambda_0\mu_0)^2}{\lambda_0+N}\right.\\
&\quad \left. + (\lambda_0+N)\left(\mu - \frac{N\bar{y}+\lambda_0\mu_0}{\lambda_0+N}\right)^2\right)\\
&= -\frac{1}{2\sigma^2}\left(2\beta_0 + \sum_{n=1}^{N}(y_n-\bar{y})^2 + \frac{(\lambda_0+N)(N\bar{y}^2+\lambda_0\mu_0^2) - N^2\bar{y}^2 - \lambda_0^2\mu_0^2 + 2N\bar{y}\lambda_0\mu_0}{\lambda_0+N}\right.\\
&\quad \left. + (\lambda_0+N)\left(\mu - \frac{N\bar{y}+\lambda_0\mu_0}{\lambda_0+N}\right)^2\right)\\
&= -\frac{1}{2\sigma^2}\left(2\beta_0 + \sum_{n=1}^{N}(y_n-\bar{y})^2 + \frac{\lambda_0 N}{\lambda_0+N}(\bar{y}-\mu_0)^2 + (\lambda_0+N)\left(\mu - \frac{N\bar{y}+\lambda_0\mu_0}{\lambda_0+N}\right)^2\right).
\end{aligned}
$$

Therefore, the above expression can be identified as a Gaussian-inverse-gamma PDF by writing:

$$
p(\mu,\sigma^2 \mid \boldsymbol{y}) \propto \left(\frac{1}{\sigma^2}\right)^{\alpha_N+\frac{3}{2}} \exp\left(-\frac{2\beta_N + \lambda_N(\mu-\mu_N)^2}{2\sigma^2}\right), \tag{5.1}
$$

with:

- $\mu_N = \dfrac{N\bar{y}+\lambda_0\mu_0}{\lambda_0+N}$,
- $\lambda_N = \lambda_0 + N$,
- $\alpha_N = \alpha_0 + \dfrac{N}{2}$,
- $\beta_N = \beta_0 + \dfrac{1}{2}\sum\limits_{n=1}^{N}(y_n-\bar{y})^2 + \dfrac{\lambda_0 N}{2(\lambda_0+N)}(\bar{y}-\mu_0)^2$.

The normalising constant is induced by this characteristic formulation and the joint posterior distribution can be expressed as:

$$
\mu,\sigma^2 \mid \boldsymbol{y} \sim \mathcal{N}\Gamma^{-1}\left(\mu_N,\lambda_N,\alpha_N,\beta_N\right) \tag{5.2}
$$

Although these update formulas provide a valuable result in itself, we shall see in the sequel that we are more interested in the marginal distribution over the mean parameter μ, for comparison purposes. Computing this marginal from the joint posterior in Equation (5.2)

remains tractable as well by integrating over σ^2:

$$
\begin{aligned}
p(\mu \mid \boldsymbol{y}) &= \int p(\mu, \sigma^2 \mid \boldsymbol{y}) \, \mathrm{d}\sigma^2 \\
&= \frac{\sqrt{\lambda_N}}{\sqrt{2\pi}} \frac{\beta_N^{\alpha_N}}{\Gamma(\alpha_N)} \int \left(\frac{1}{\sigma^2}\right)^{\alpha_N + \frac{3}{2}} \exp\left(-\frac{2\beta_N + \lambda_N(\mu - \mu_N)^2}{2\sigma^2}\right) \mathrm{d}\sigma^2 \\
&= \frac{\sqrt{\lambda_N}}{\sqrt{2\pi}} \frac{\beta_N^{\alpha_N}}{\Gamma(\alpha_N)} \frac{\Gamma(\alpha_N + \frac{1}{2})}{(\beta_N + \frac{\lambda_N}{2}(\mu - \mu_N)^2)^{\alpha_N + \frac{1}{2}}} \\
&\quad \times \int \underbrace{\frac{(\beta_N + \frac{\lambda_N}{2}(\mu - \mu_N)^2)^{\alpha_N + \frac{1}{2}}}{\Gamma(\alpha_N + \frac{1}{2})} \left(\frac{1}{\sigma^2}\right)^{\alpha_N + \frac{1}{2} + 1} \exp\left(-\frac{\beta_N + \frac{\lambda_N}{2}(\mu - \mu_N)^2}{\sigma^2}\right)}_{\Gamma^{-1}(\alpha_N + \frac{1}{2},\ \beta_N + \frac{\lambda_N}{2}(\mu - \mu_N)^2)} \mathrm{d}\sigma^2 \\
&= \frac{\sqrt{\lambda_N}}{\sqrt{2\pi}} \frac{\beta_N^{\alpha_N}}{\Gamma(\alpha_N)} \frac{\Gamma(\alpha_N + \frac{1}{2})}{(\beta_N + \frac{\lambda_N}{2}(\mu - \mu_N)^2)^{\alpha_N + \frac{1}{2}}} \times 1 \\
&= \frac{\Gamma(\alpha_N + \frac{1}{2})}{\Gamma(\alpha_N)} \frac{\sqrt{\lambda_N}}{\sqrt{2\pi}} \frac{\beta_N^{\alpha_N + \frac{1}{2}}}{\sqrt{\beta_N}} (\beta_N + \frac{\lambda_N}{2}(\mu - \mu_N)^2)^{-\alpha_N - \frac{1}{2}} \\
&= \frac{\Gamma(\alpha_N + \frac{1}{2})}{\Gamma(\alpha_N)} \frac{\sqrt{\lambda_N}}{\sqrt{2\pi}} \frac{\beta_N^{\alpha_N + \frac{1}{2}}}{\sqrt{\beta_N}} \beta_N^{-\alpha_N - \frac{1}{2}} (1 + \frac{\alpha_N \lambda_N}{2\alpha_N \beta_N}(\mu - \mu_N)^2)^{-\alpha_N - \frac{1}{2}} \\
&= \frac{\Gamma(\alpha_N + \frac{1}{2})}{\Gamma(\alpha_N)} \frac{\sqrt{\alpha_N \lambda_N}}{\sqrt{2\alpha_N \pi \beta_N}} (1 + \frac{1}{2\alpha_N} \frac{\alpha_N \lambda_N (\mu - \mu_N)^2}{\beta_N})^{-\alpha_N - \frac{1}{2}} \\
&= \underbrace{\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\pi \nu \hat{\sigma}^2}} (1 + \frac{1}{\nu} \frac{(\mu - \mu_N)^2}{\hat{\sigma}^2})^{-\frac{\nu+1}{2}}}_{T_\nu(\mu;\ \mu_N, \hat{\sigma}^2)},
\end{aligned}
$$

with:

- $\nu = 2\alpha_N$,
- $\hat{\sigma}^2 = \dfrac{\beta_N}{\alpha_N \lambda_N}$.

The marginal posterior distribution over μ can thus be expressed as a non-standardised Student's t-distribution that we express below in terms of the initial hyper-parameters:

$$
\mu \mid \boldsymbol{y} \sim T_{2\alpha_0 + N} \left(\frac{N\bar{y} + \lambda_0 \mu_0}{\lambda_0 + N}, \frac{\beta_0 + \frac{1}{2} \sum_{n=1}^{N} (y_n - \bar{y})^2 + \frac{\lambda_0 N}{2(\lambda_0 + N)} (\bar{y} - \mu_0)^2}{(\alpha_0 + \frac{N}{2})(\lambda_0 + N)} \right). \tag{5.3}
$$

The derivation of this analytical formula provides a valuable tool for computing straightforward posterior distribution for the mean parameter in such context. We shall see in the next section how to leverage this approach to introduce a novel means' comparison methodology for a more general framework, to handle both multidimensional and missing data.

5.2 General Bayesian framework for evaluating mean differences ...

Recalling our differential proteomics context that consists in assessing the differences in mean intensity values for P peptides or proteins quantified in N samples divided into K conditions. As before, Figure 5.2 illustrates the hierarchical generative structure assumed for each group $k = 1, \dots, K$.

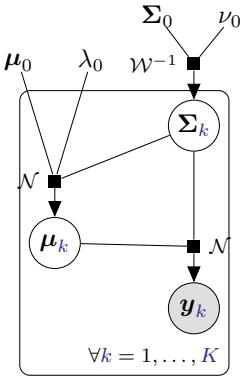


Figure 5.2: Graphical model of the hierarchical structure of the generative model for the vector $\boldsymbol{y}_k$ of peptide intensities in K groups of biological samples, *i.e.* K experimental conditions.

Maintaining the notation analogous to previous ones, the generative model for $\boldsymbol{y}_k \in \mathbb{R}^P$, can be written as:

$$\boldsymbol{y}_k = \boldsymbol{\mu}_k + \boldsymbol{\varepsilon}_k, \ \forall k = 1, \dots, K,$$

where:

- $\boldsymbol{\mu}_k \mid \boldsymbol{\Sigma}_k \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{1}{\lambda_0}\boldsymbol{\Sigma}_k\right)$ is the prior mean intensities vector of the k-th group,
- $\boldsymbol{\varepsilon}_k \sim \mathcal{N}(0, \boldsymbol{\Sigma}_k)$ is the error term of the k-th group,
- $\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(\boldsymbol{\Sigma}_0, \nu_0)$ is the prior variance-covariance matrix of the k-th group,

with $\{\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Sigma}_0, \nu_0\}$ a set of hyper-parameters that needs to be chosen as modelling hypotheses and $\mathcal{W}^{-1}$ represents the Inverse-Wishart distribution, previously introduced in **??**, and used as the conjugate prior for an unknown covariance matrix of a multivariate Gaussian distribution.

Traditionally, in Bayesian inference, those quantities need to be carefully chosen for the estimation to be as accurate as possible, in particular with low sample sizes. The incorporation of expert or prior knowledge on the model would also come from the adequate setting of these hyper-parameters. However, our final purpose in this chapter is not much about estimating but instead focused on comparing groups' mean (i.e. differential analysis). Interestingly, providing a perfect estimation of the posterior distributions over

$\{\boldsymbol{\mu}_k\}_{k=1,\dots,K}$ does not appear as the main concern here, as the posterior difference of means (i.e. $p(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'} \mid \boldsymbol{y}_k, \boldsymbol{y}_{k'})$) represents the actual quantity of interest. Although providing meaningful prior hyper-parameters leads to adequate uncertainty quantification, we shall, above all, take those quantities equal for all groups. This choice would ensure an unbiased comparison, which would constitute a valuable alternative to the traditional and somehow limited t-tests. Indeed, inference based on hypothesis testing and p-values has been widely called into question over the past decade (Wasserstein et al., 2019). Additionally, t-tests do not provide any insight on effect sizes or uncertainty quantification (in contrast to Bayesian inference as emphasized by Kruschke and Liddell (2018)).

The present framework aspires at estimating a posterior distribution for each mean parameter vector $\boldsymbol{\mu}_k$, starting from the same prior assumptions in each group. The comparison between means of all groups would then only rely on the ability to sample directly from these distributions and compute empirical posteriors for the means' difference. As a bonus, this framework remains compatible with multiple imputations strategies previously introduced to handle missing data that frequently arise in applicative contexts (see Chapter 3).

From the previous hypotheses, we can deduce the likelihood of the model for an i.i.d. sample $\{\boldsymbol{y}_{k,1}, \dots, \boldsymbol{y}_{k,N_k}\}$:

$$\begin{aligned} p(\boldsymbol{y}_{k,1}, \dots, \boldsymbol{y}_{k,N_k} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= \prod_{n=1}^{N_k} p(\boldsymbol{y}_{k,n} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ &= \prod_{n=1}^{N_k} \mathcal{N}\left(\boldsymbol{y}_{k,n};\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \end{aligned}$$

However, as previously pointed out, such datasets often contain missing data and we shall introduce here consistent notation. Assume $\mathcal{H}$ to be the set of all observed data, we additionally define:

- $\boldsymbol{y}_k^{(0)} = \{y_{k,n}^p \in \mathcal{H},\ n = 1, \dots N_k,\ p = 1, \dots, P\}$, the set of elements that are observed in the k-th group,
- $\boldsymbol{y}_k^{(1)} = \{y_{k,n}^p \notin \mathcal{H},\ n = 1, \dots N_k,\ p = 1, \dots, P\}$, the set of elements that are missing the k-th group.

Moreover, as we remain in the context of multiple imputation, $\{\tilde{\boldsymbol{y}}_k^{(1),1}, \dots, \tilde{\boldsymbol{y}}_k^{(1),D}\}$ can be defined as the set of D draws of an imputation process applied on missing data in the k-th group. In such context, a closed-form approximation for the multiple-imputed posterior distribution of $\boldsymbol{\mu}_k$ can be derived for each group as stated in Proposition 5.1.

Proposition 5.1. *For all $k = 1, \dots, K$, the posterior distribution of $\boldsymbol{\mu}_k$ can be approximated*

by a mixture of multiple-imputed multivariate t-distributions, such as:

$$p(\boldsymbol{\mu}_k \mid \boldsymbol{y}_k^{(0)}) \simeq \frac{1}{D}\sum_{d=1}^{D} T_{\nu_k}\left(\boldsymbol{\mu}; \tilde{\boldsymbol{\mu}}_k^{(d)}, \tilde{\boldsymbol{\Sigma}}_k^{(d)}\right)$$

with:

- $\nu_k = \nu_0 + N_k - P + 1$,
- $\tilde{\boldsymbol{\mu}}_k^{(d)} = \dfrac{\lambda_0\boldsymbol{\mu}_0 + N_k\bar{\boldsymbol{y}}_k^{(d)}}{\lambda_0 + N_k}$,
- $\tilde{\boldsymbol{\Sigma}}_k^{(d)} = \dfrac{\boldsymbol{\Sigma}_0 + \sum_{n=1}^{N_k}(\tilde{\boldsymbol{y}}_{k,n}^{(d)} - \bar{\boldsymbol{y}}_k^{(d)})(\tilde{\boldsymbol{y}}_{k,n}^{(d)} - \bar{\boldsymbol{y}}_k^{(d)})^\intercal + \dfrac{\lambda_0 N_k}{(\lambda_0 + N_k)}(\bar{\boldsymbol{y}}_k^{(d)} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{y}}_k^{(d)} - \boldsymbol{\mu}_0)^\intercal}{(\nu_0 + N_k - P + 1)(\lambda_0 + N_k)}$,

where we introduced the shorthand $\tilde{\boldsymbol{y}}_{k,n}^{(d)} = \begin{bmatrix} \boldsymbol{y}_{k,n}^{(0)} \\ \tilde{\boldsymbol{y}}_{k,n}^{(1),d} \end{bmatrix}$ *to represent the* d*-th imputed vector of observed data, and the corresponding average vector* $\bar{\boldsymbol{y}}_k^{(d)} = \dfrac{1}{N_k}\sum_{n=1}^{N_k}\tilde{\boldsymbol{y}}_{k,n}^{(d)}$.

This analytical formulation is particularly convenient for our purpose and, as we shall see in the proof below, merely comes from imputation.

Proof. For the sake of clarity, let us omit the k groups here and first consider a general case with $\boldsymbol{y}_k = \boldsymbol{y} \in \mathbb{R}^P$. Moreover, let us focus on only one imputed dataset, and maintain the notation $\tilde{\boldsymbol{y}}_1^{(d)}, \dots, \tilde{\boldsymbol{y}}_N^{(d)} = \boldsymbol{y}_1, \dots, \boldsymbol{y}_N$ for convenience. From the hypotheses of the model, we can derive $\mathcal{L}$, the posterior log-PDF over $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, following the same idea as for the univariate case presented Section 5.1:

$$\begin{aligned}
\mathcal{L} &= \log p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_N) \\
&= \log \underbrace{p(\boldsymbol{y}_1, \dots, \boldsymbol{y}_N \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})}_{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} + \log \underbrace{p(\boldsymbol{\mu}, \boldsymbol{\Sigma})}_{\mathcal{NW}^{-1}(\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Sigma}_0, \nu_0)} + C_1 \\
&= -\frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\left(\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu})\right) \\
&\quad - \frac{\nu_0 + P + 2}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\left(\operatorname{tr}\left(\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}\right) - \frac{\lambda_0}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right) + C_2 \\
&= -\frac{1}{2}\Bigg[(\nu_0 + P + 2 + N)\log|\boldsymbol{\Sigma}| + \operatorname{tr}\left(\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}\right) \\
&\quad + \sum_{n=1}^{N}\operatorname{tr}\left((\boldsymbol{y}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu})\right) + \operatorname{tr}\left(\lambda_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right)\Bigg] + C_2
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2}\Bigg[(\nu_0 + P + 2 + N)\log|\boldsymbol{\Sigma}| + \operatorname{tr}\Bigg(\boldsymbol{\Sigma}^{-1}\Big\{ \boldsymbol{\Sigma}_0 + \lambda_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0)(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\intercal \\
&\quad \underbrace{+ N(\bar{\boldsymbol{y}}-\boldsymbol{\mu})(\bar{\boldsymbol{y}}-\boldsymbol{\mu})^\intercal + \sum_{n=1}^{N}(\boldsymbol{y}_n-\bar{\boldsymbol{y}})(\boldsymbol{y}_n-\bar{\boldsymbol{y}})^\intercal}_{\text{Lemma 1}} \Big\}\Bigg)\Bigg] + C_2 \\
&= -\frac{1}{2}\Bigg[(\nu_0 + P + 2 + N)\log|\boldsymbol{\Sigma}| + \operatorname{tr}\Bigg(\boldsymbol{\Sigma}^{-1}\Big\{ \boldsymbol{\Sigma}_0 + \sum_{n=1}^{N}(\boldsymbol{y}_n-\bar{\boldsymbol{y}})(\boldsymbol{y}_n-\bar{\boldsymbol{y}})^\intercal \\
&\quad + (N+\lambda_0)\boldsymbol{\mu}\boldsymbol{\mu}^\intercal - \boldsymbol{\mu}\left(N\bar{\boldsymbol{y}}^\intercal + \lambda_0\boldsymbol{\mu}_0^\intercal\right) - \left(\lambda_0\boldsymbol{\mu}_0 + N\bar{\boldsymbol{y}}\right)\boldsymbol{\mu}^\intercal + \lambda_0\boldsymbol{\mu}_0\boldsymbol{\mu}_0^\intercal + N\bar{\boldsymbol{y}}\bar{\boldsymbol{y}}^\intercal \Big\}\Bigg)\Bigg] + C_2 \\
&= -\frac{1}{2}\Bigg[(\nu_0 + P + 2 + N)\log|\boldsymbol{\Sigma}| \\
&\quad + \operatorname{tr}\Bigg(\boldsymbol{\Sigma}^{-1}\Big\{ \boldsymbol{\Sigma}_0 + \sum_{n=1}^{N}(\boldsymbol{y}_n-\bar{\boldsymbol{y}})(\boldsymbol{y}_n-\bar{\boldsymbol{y}})^\intercal + \frac{N\lambda_0}{N+\lambda_0}(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)^\intercal \\
&\quad + (N+\lambda_0)\left(\boldsymbol{\mu} - \frac{N\bar{\boldsymbol{y}}+\lambda_0\boldsymbol{\mu}_0}{N+\lambda_0}\right)\left(\boldsymbol{\mu} - \frac{N\bar{\boldsymbol{y}}+\lambda_0\boldsymbol{\mu}_0}{N+\lambda_0}\right)^\intercal \Big\}\Bigg)\Bigg] + C_2 \\
&= -\frac{1}{2}\Bigg[(\nu_N + P + 2)\log|\boldsymbol{\Sigma}| + \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_N\right) + \lambda_N\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)^\intercal\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)\Bigg] + C_2.
\end{aligned}
$$

By identification, we recognise the log-PDF that characterises the Gaussian-inverse-Wishart distribution $\mathcal{NIW}^{-1}(\boldsymbol{\mu}_N, \lambda_N, \boldsymbol{\Sigma}_N, \nu_N)$ with:

- $\boldsymbol{\mu}_N = \dfrac{N\bar{\boldsymbol{y}} + \lambda_0\boldsymbol{\mu}_0}{N+\lambda_0}$,
- $\lambda_N = \lambda_0 + N$,
- $\boldsymbol{\Sigma}_N = \boldsymbol{\Sigma}_0 + \sum\limits_{n=1}^{N}(\boldsymbol{y}_N-\bar{\boldsymbol{y}})(\boldsymbol{y}_N-\bar{\boldsymbol{y}})^\intercal + \dfrac{\lambda_0 N}{(\lambda_0+N)}(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)^\intercal$,
- $\nu_N = \nu_0 + N$.

Once more, we can integrate over $\boldsymbol{\Sigma}$ to compute the mean's marginal posterior distribution by identifying the PDF of the inverse-Wishart distribution $\mathcal{W}^{-1}\Big(\boldsymbol{\Sigma}_N + \lambda_N\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)^\intercal, \nu_N + 1\Big)$ and by reorganising the terms:

$$
\begin{aligned}
p(\boldsymbol{\mu} \mid \boldsymbol{y}) &= \int p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{y})\,\mathrm{d}\boldsymbol{\Sigma} \\
&= \frac{\lambda_N^{\frac{P}{2}}|\boldsymbol{\Sigma}_N|^{\frac{\nu_N}{2}}}{(2\pi)^{\frac{P}{2}}2^{\frac{P\nu_N}{2}}\Gamma_P\left(\frac{\nu_N}{2}\right)} \\
&\quad \times \int |\boldsymbol{\Sigma}|^{-\frac{\nu_N+P+2}{2}} \exp\left(-\frac{1}{2}\left(\operatorname{tr}\left(\boldsymbol{\Sigma}_N\boldsymbol{\Sigma}^{-1}\right) - \frac{\lambda_N}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)^\intercal\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)\right)\right)\mathrm{d}\boldsymbol{\Sigma}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\lambda_N^{\frac{P}{2}} |\boldsymbol{\Sigma}_{\boldsymbol{N}}|^{\frac{\nu_N}{2}}}{(2\pi)^{\frac{P}{2}} 2^{\frac{P\nu_N}{2}} \Gamma_P\left(\frac{\nu_N}{2}\right)} \times \frac{2^{\frac{P(\nu_N+1)}{2}} \Gamma_P\left(\frac{\nu_N+1}{2}\right)}{|\boldsymbol{\Sigma}_N + \lambda_N (\boldsymbol{\mu} - \boldsymbol{\mu}_N)(\boldsymbol{\mu} - \boldsymbol{\mu}_N)^\intercal|^{\frac{\nu_N+1}{2}}} \times 1 \\
&= \frac{\pi^{p(p-1)/4} \prod_{p=0}^{P-1} \Gamma\left(\frac{\nu_N+1-p}{2}\right)}{\pi^{P(P-1)/4} \prod_{p=1}^{P} \Gamma\left(\frac{\nu_N+1-p}{2}\right)} \times \frac{\lambda_N^{\frac{P}{2}}}{\pi^{\frac{P}{2}}} \\
&\quad \times \underbrace{\frac{|\boldsymbol{\Sigma}_N|^{\frac{\nu_N}{2}}}{|\boldsymbol{\Sigma}_N|^{\frac{\nu_N+1}{2}}} \times \left(1 + \lambda_N (\boldsymbol{\mu} - \boldsymbol{\mu}_N)^\intercal \boldsymbol{\Sigma}_N^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_N)\right)^{-\frac{\nu_N+1}{2}}}_{\text{Matrix determinant lemma}} \\
&= \frac{\Gamma\left(\frac{\nu_N+1}{2}\right)}{\Gamma\left(\frac{\nu_N+1-P}{2}\right)} \times \frac{[\lambda_N(\nu_N - P + 1)]^{\frac{P}{2}}}{[\pi(\nu_N - P + 1)]^{\frac{P}{2}} |\boldsymbol{\Sigma}_N|^{\frac{1}{2}}} \\
&\quad \times \left(1 + \frac{\lambda_N(\nu_N - P + 1)}{(\nu_N - P + 1)} (\boldsymbol{\mu} - \boldsymbol{\mu}_N)^\intercal \boldsymbol{\Sigma}_N^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_N)\right)^{-\frac{\nu_N+1}{2}} \\
&= \frac{\Gamma\left(\frac{(\nu_N-P+1)+P}{2}\right)}{\Gamma\left(\frac{\nu_N-P+1}{2}\right) [\pi(\nu_N - P + 1)]^{\frac{P}{2}} |\frac{\boldsymbol{\Sigma}_N}{\lambda_N(\nu_N - P + 1)}|^{\frac{1}{2}}} \\
&\quad \times \left(1 + \frac{1}{\nu_N - P + 1} (\boldsymbol{\mu} - \boldsymbol{\mu}_N)^\intercal \left(\frac{\boldsymbol{\Sigma}_N}{\lambda_N(\nu_N - P + 1)}\right)^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_N)\right)^{-\frac{(\nu_N-P+1)+P}{2}}.
\end{aligned}
$$

The above expression corresponds to the PDF of a multivariate t-distribution $\mathcal{T}_\nu\left(\boldsymbol{\mu}_N, \hat{\boldsymbol{\Sigma}}\right)$, with:

- $\nu = \nu_N - P + 1$,
- $\hat{\boldsymbol{\Sigma}} = \dfrac{\boldsymbol{\Sigma}_N}{\lambda_N(\nu_N - P + 1)}$.

Therefore, we demonstrated that for each group and imputed dataset, the complete-data posterior over $\boldsymbol{\mu}_k$ happens to be a multivariate t-distribution. Thus, following Rubin's rules for multiple imputation (see Equation (1.19) in Section 1.2.3.b), we can propose an approximation to the true posterior distribution (that is only conditioned over observed values):

$$
\begin{aligned}
p\left(\boldsymbol{\mu}_k \mid \boldsymbol{y}_k^{(0)}\right) &= \int p\left(\boldsymbol{\mu}_k \mid \boldsymbol{y}_k^{(0)}, \boldsymbol{y}_k^{(1)}\right) p\left(\boldsymbol{y}_k^{(1)} \mid \boldsymbol{y}_k^{(0)}\right) \mathrm{d}\boldsymbol{y}_k^{(1)} \\
&\simeq \frac{1}{P} \sum_{p=1}^{P} p\left(\boldsymbol{\mu}_k \mid \boldsymbol{y}_k^{(0)}, \tilde{\boldsymbol{y}}_k^{(1),d}\right)
\end{aligned}
$$

Leading to the desired results when evaluating the previously derived posterior distribution on each multiple-imputed dataset. □

Thanks to Proposition 5.1, we have an explicit formula for approximating, using multiple

imputed datasets, the posterior distribution of the mean vector for each group. Although such linear combination of multivariate t-distributions is not a known specific distribution in itself, it is now straightforward to generate realisations of samples of the posterior by simply drawing from the D multivariate t-distributions, each being specific to an imputed dataset, and then compute the mean of the D vectors. Therefore, the empirical distribution resulting from a high number of samples generated by this procedure would be easy to visualise and manage for comparison purpose. Generating the empirical distribution of the mean's difference between two groups k and k' then comes directly, by computing the difference between each couple of samples drawn from both posterior distributions $p(\boldsymbol{\mu}_k \mid \boldsymbol{y}_k^{(0)})$ and $p(\boldsymbol{\mu}'_k \mid \boldsymbol{y}_{k'}^{(0)})$. In Bayesian statistics, relying on empirical distributions drawn from the posterior is common practice in the context of Markov chain Monte Carlo (MCMC) algorithms, but often comes at a high computational cost. In our framework, we managed to maintain the best of both worlds since deriving analytical distributions from model hypotheses offers the benefits of probabilistic inference with adequate uncertainty quantification, while remaining tractable and not relying on MCMC procedures. The computational cost of the method thus roughly remains as low as frequentist counterparts since merely a few updating calculus and drawing from t-distributions are needed.

As usual when it comes to compare the mean between two groups, we still need to assess if the posterior distribution of the difference appear, in a sense, to be sufficiently away from zero. This practical inference choice is not specific to our context and remains highly dependent on the context of the study. Moreover, as the present model is multi-dimensional, we may also raise the question of the metric used to compute the difference between vectors. In a sense, our posterior distribution of the mean's differences offers an elegant solution to the traditional problem of multiple testing often encountered in applied science and allows tailored definitions of what could be called a *meaningful* result (*significant* does not appear anymore as an appropriate term in this more general context). For example, displaying the distribution of the squared difference would penalise large differences in elements of the mean vector whereas absolute difference would give a more balanced conception of the average divergence from one group to the other. Clearly, as any marginal of a multivariate t-distribution remains a (multivariate) t-distribution, it is also straightforward to compare specific elements of the mean vectors merely by restraining to the appropriate dimension. Recalling our proteomics context, this means that we could still compare mean intensity of peptides between groups one peptide at a time, or choosing to compare all peptides at once and thus accounting for possible correlations between peptides in each group. However, an appropriate manner to account for those correlations could be to subset peptides using their protein groups.

Let us provide in Algorithm 1 a summary of the whole procedure for comparing mean vectors of two different experimental conditions in terms of posterior distribution.

Algorithm 1 Posterior distribution of the vector of mean's difference

Initialise the hyper-posteriors $\boldsymbol{\mu}_0^k = \boldsymbol{\mu}_0^{k'}$, $\lambda_0^k = \lambda_0^{k'}$, $\boldsymbol{\Sigma}_0^k = \boldsymbol{\Sigma}_0^{k'}$, $\nu_0^k = \nu_0^{k'}$

for $d = 1, \ldots, D$ **do**

Compute $\{\boldsymbol{\mu}_N^{k,(d)}, \lambda_N^k, \boldsymbol{\Sigma}_N^{k,(d)}, \nu_N^k\}$ and $\{\boldsymbol{\mu}_N^{k',(d)}, \lambda_N^{k'}, \boldsymbol{\Sigma}_N^{k',(d)}, \nu_N^{k'}\}$ from hyper-posteriors and data

Draw R realisations $\hat{\boldsymbol{\mu}}_k^{(d)[r]} \sim T_{\nu_N^k}\left(\boldsymbol{\mu}_N^{k,(d)}, \dfrac{\boldsymbol{\Sigma}_N^{k,(d)}}{\lambda_N^k \nu_N^k}\right)$; $\hat{\boldsymbol{\mu}}_{k'}^{(d)[r]} \sim T_{\nu_N^{k'}}\left(\boldsymbol{\mu}_N^{k',(d)}, \dfrac{\boldsymbol{\Sigma}_N^{k',(d)}}{\lambda_N^{k'} \nu_N^{k'}}\right)$

end for

for $r = 1, \ldots, R$ **do**

Compute $\hat{\boldsymbol{\mu}}_k^{[r]} = \dfrac{1}{D}\sum_{d=1}^{D} \hat{\boldsymbol{\mu}}_k^{(d)[s]}$ and $\hat{\boldsymbol{\mu}}_{k'}^{[r]} = \dfrac{1}{D}\sum_{d=1}^{D} \hat{\boldsymbol{\mu}}_{k'}^{(d)[r]}$ to combine samples

Generate a realisation $\hat{\boldsymbol{\mu}}_\Delta^{[r]} = \hat{\boldsymbol{\mu}}_k^{[r]} - \hat{\boldsymbol{\mu}}_{k'}^{[r]}$ from the difference's distribution

end for

return $\{\hat{\boldsymbol{\mu}}_\Delta^{[1]}, \ldots, \hat{\boldsymbol{\mu}}_\Delta^{[R]}\}$, an R-sample drawn from the posterior distribution of the mean's difference

5.3 The uncorrelated case: no more multiple testing nor imputation

Let us notice that modelling covariances between all variables as in Proposition 5.1 often constitutes a challenge, which is computationally expensive in high dimensions and not always adapted. However, we detailed in Section 5.1 results that are classical in Bayesian inference, but somehow not widespread enough in applied science, especially when it comes to comparing means. In particular, we can leverage these results to adapt Algorithm 1 to the univariate case, for handling the same problem as in Chapter 3 with a more probabilistic flavour. Indeed, when the absence of correlations between peptides is assumed (*i.e.* $\boldsymbol{\Sigma}$ being diagonal), the problem reduces to the analysis of P independent inference problems (as $\boldsymbol{\mu}$ is supposed Gaussian) and the posterior distributions can be derived in closed-form, as we recalled in Equation (5.1). Moreover, let us highlight a nice property coming with this relaxing assumption is that (multiple-)imputation is no longer needed in this context. Using the same notation as before and the uncorrelated assumption (and thus the induced independence between analytes for $p \neq p'$), we can write:

$$p\left(\boldsymbol{\mu}_k \mid \boldsymbol{y}_k^{(0)}\right) = \int p\left(\boldsymbol{\mu}_k, \boldsymbol{y}_k^{(1)} \mid \boldsymbol{y}_k^{(0)}\right) \mathrm{d}\boldsymbol{y}_k^{(1)} \tag{5.4}$$

$$= \int p\left(\boldsymbol{\mu}_k \mid \boldsymbol{y}_k^{(0)}, \boldsymbol{y}_k^{(1)}\right) p\left(\boldsymbol{y}_k^{(1)} \mid \boldsymbol{y}_k^{(0)}\right) \mathrm{d}\boldsymbol{y}_k^{(1)} \tag{5.5}$$

$$= \int \prod_{p=1}^{P} \left\{ p\left(\mu_k^p \mid y_k^{p,(0)}, y_k^{p,(1)}\right) p\left(y_k^{p,(1)} \mid y_k^{p,(0)}\right) \right\} \mathrm{d}\boldsymbol{y}_k^{(1)} \tag{5.6}$$

$$= \prod_{p=1}^{P} \int \left\{ p\left(\mu_k^p \mid y_k^{p,(0)}, y_k^{p,(1)}\right) p\left(y_k^{p,(1)} \mid y_k^{p,(0)}\right) \mathrm{d}y_k^{p,(1)} \right\} \tag{5.7}$$

$$= \prod_{p=1}^{P} p\left(\mu_k^p \mid y_k^{p,(0)}\right) \tag{5.8}$$

$$= \prod_{p=1}^{P} T_{2\alpha_0^p + N_k^p}\left(\mu_k^p;\ \mu_{k,N}^p,\ \hat{\sigma_k^p}^2\right), \tag{5.9}$$

with:

- $\mu_{k,N}^p = \dfrac{N_k^p \bar{y}_k^{p,(0)} + \lambda_0^p \mu_0^p}{\lambda_0^p + N_k^p}$,
- $\hat{\sigma_k^p}^2 = \dfrac{\beta_0^p + \frac{1}{2}\sum_{n=1}^{N_k^p} (y_{k,n}^{p,(0)} - \bar{y}_k^{p,(0)})^2 + \frac{\lambda_0 N_k^p}{2(\lambda_0^p + N_k^p)} (\bar{y}_k^{p,(0)} - \mu_0^p)^2}{(\alpha_0^p + \frac{N_k^p}{2})(\lambda_0^p + N_k^p)}$.

In this context, it can be noticed that $p\left(\boldsymbol{\mu}_k \mid \boldsymbol{y}_k^{(0)}\right)$ factorises naturally over $p = 1, \ldots, P$, and thus only depends upon the data that have actually been observed for each peptide. Indeed, we observe that the integration over the missing data $\boldsymbol{y}_k^{(1)}$ is straightforward in this framework and neither the Rubin's approximation or even imputation (whether multiple or not) appear necessary. The observed data $\boldsymbol{y}_k^{(0)}$ already bear all the useful information as if each unobserved values could simply be ignored without effect on the posterior distribution.

Let us emphasise on the fact that this property of factorisation and tractable integration over missing data comes directly from the covariance structure as a diagonal matrix, and thus only constitutes a particular case, though convenient, of the previous model. However, in the context of differential analysis in proteomics, analysing each peptide as an independent problem is a common practice, as seen in Chapter 3, and we shall notice that the Bayesian framework tackles this issue in an elegant and somehow simpler way. In particular, the classical inference approach based on hypothesis testing performs numerous successive tests for all peptides. Such an approach often leads to the pitfall of multiple testing which needs to be carefully dealt with. Interestingly, we can notice that the above model also avoid multiple testing (as it does not rely on hypothesis testing and the definition of some threshold) while maintaining the convenient interpretations of Bayesian probabilistic inference. To conclude, whereas the analytical derivation of posterior distributions with Gaussian-inverse-gamma constitutes a well-known results, our proposition to define such probabilistic mean's comparison procedure provides, under the standard uncorrelated-peptides assumption, an

elegant and handy alternative to classical techniques that naturally tackles both the imputation and multiple testing issues. Let us provide in Algorithm 2 the pseudo-code of the inference procedure in order to highlight differences with the fully-correlated case:

Algorithm 2 Posterior distribution of the mean's difference

for $p = 1, \ldots, P$ **do**

Initialise the hyper-posteriors $\mu_0^{k,p} = \mu_0^{k',p}$, $\lambda_0^{k,p} = \lambda_0^{k',p}$, $\alpha_0^{k,p} = \alpha_0^{k',p}$, $\beta_0^{k,p} = \beta_0^{k',p}$

Compute $\{\mu_N^{k,p}, \lambda_N^{k,p}, \alpha_N^{k,p}, \beta_N^{k,p}\}$ and $\{\mu_N^{k',p}, \lambda_N^{k',p}, \alpha_N^{k',p}, \beta_N^{k',p}\}$ from hyper-posteriors and data

Draw R realisations $\hat{\mu}_k^{p,[r]} \sim T_{\alpha_N^{k,p}}\left(\mu_N^{k,p}, \frac{\beta_N^{k,p}}{\lambda_N^{k,p}\alpha_N^{k,p}}\right)$, $\hat{\mu}_{k'}^{p,[r]} \sim T_{\alpha_N^{k',p}}\left(\mu_N^{k',p}, \frac{\beta_N^{k',p}}{\lambda_N^{k',p}\alpha_N^{k',p}}\right)$

for $r = 1, \ldots, R$ **do**

Generate a realisation $\hat{\mu}_\Delta^{p,[r]} = \hat{\mu}_k^{p,[r]} - \hat{\mu}_{k'}^{p,[r]}$ from the difference's distribution

end for

end for

return $\{\hat{\boldsymbol{\mu}}_\Delta^{[1]}, \ldots, \hat{\boldsymbol{\mu}}_\Delta^{[R]}\}$, an R-sample drawn from the posterior distribution of the mean's difference

5.4 Experiments

To illustrate our methodology, we used a real proteomics dataset already introduced in Chapter 3, namely the *Arabidopsis thaliana* + UPS dataset, with the Match between Runs algorithm and at least one quantified value in each experimental condition. Briefly, let us recall that UPS proteins were spiked in increasing amounts into a constant background of *Arabidopsis thaliana* (ARATH) protein lysate. Hence, UPS proteins are differentially expressed, and ARATH proteins are not. For illustration purposes, we arbitrarily chose to focus the examples on the `P12081ups|SYHC_HUMAN_UPS` and the `sp|F4I893|ILA_ARATH` proteins. Note that both proteins have nine quantified peptides. Unless otherwise stated, we took the examples of the `AALEELVK` UPS peptide and the `VLPLIIPILSK` ARATH peptide and the following values have been set for the prior hyper-parameters:

- $\mu_0 = 20, \ \forall p = 1, \ldots, P$,
- $\lambda = 1$,
- $\alpha_0 = 1$,
- $\beta_0 = 1$,
- $\boldsymbol{\Sigma}_0 = I_P$,

- $\nu_0 = 10$.

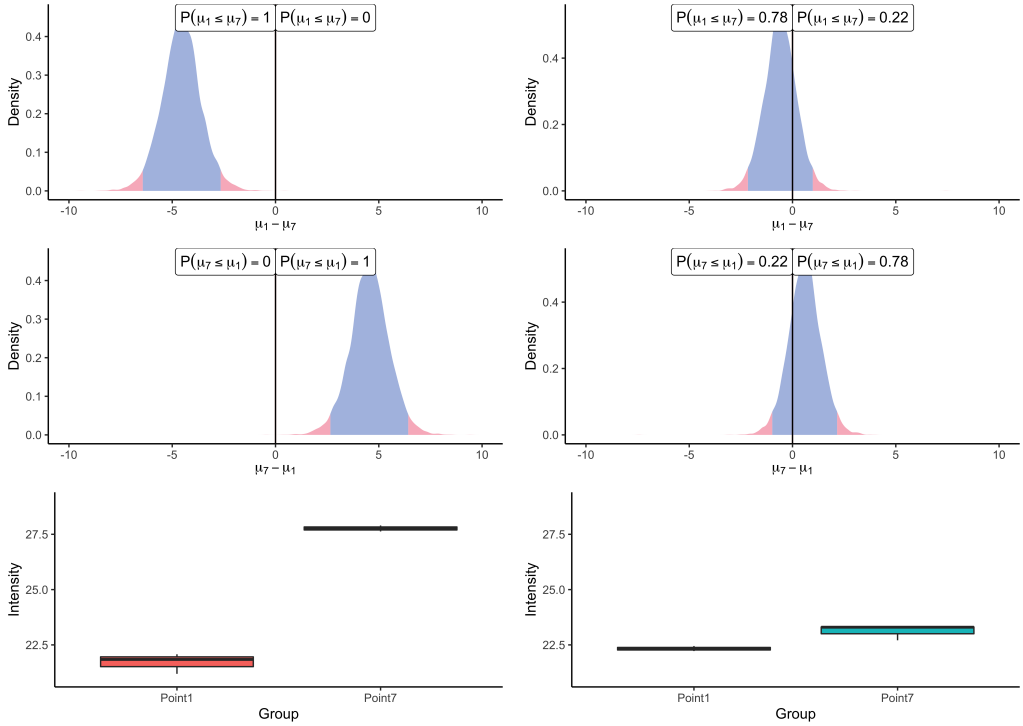
These values correspond to the practical insights acquired from our previous studies, while remaining relatively vague in terms of prior variance. As previously stated, it is essential for these values to be identical in all groups for ensuring a fair and unbiased comparison. In the case where more expert information would be accessible, its incorporation would be possible, for instance, through the definition of a more precise prior mean (μ_0) associated with a more confident prior variance (encoded through α_0 and β_0). Additionally let us recall that in our real datasets, the constants of the values take the values:

- $\forall k = 1, \ldots, K,\ N_k = 3$ data points, in the absence of missing data,
- $P = 9$ peptides, when using the multivariate model,
- $D = 7$ draws of imputation,
- $R = 10^4$ sample points from the posterior distributions.

Let us emphasise that, in this context where the number N_k of observed biological samples is extremely low, in particular when data are missing, we should expect a perceptible influence of the prior hyper-parameters, as well as an inherent uncertainty in the posteriors. However, this influence has been reduced to the minimum in all the subsequent graphs for the sake of clarity and for assuring a good understanding of the underlying properties of the methodology. The high number R of sample points drawn from the posteriors assures the empirical distribution to be smoothly displayed on the graph, but one should note that sampling is really quick in practice, and this number can be easily increased if necessary.

5.4.1 Univariate Bayesian inference for differential analysis

First, let us illustrate the univariate framework described in Section 5.3. In this experience, we compared the intensity means in the lowest (0.05 fmol UPS) and the highest points (10 fmol UPS) of the UPS spike range. Let us recall that our univariate algorithm does not rely on imputation and should be applied directly on raw data. For the sake of illustration, the chosen peptides were observed entirely in all three biological samples of both experimental conditions. Resulting from the application of our univariate algorithm, posterior distributions of the mean difference for both peptides are represented on Figure 5.3. As the analysis consists in a comparison between conditions, the 0 value has been highlighted on the x-axis for assessing both the direction and the magnitude of the difference. The distance to zero of the distributions indicates whether the peptide is differentially expressed or not. In particular, Figure 5.3a shows the posterior distribution of the means difference for the UPS peptide. Its location, far from zero, indicates a high probability (almost surely in this case) that the mean intensity of this peptide differs between the two considered groups. Conversely, the posterior distribution of the difference of means for the ARATH peptide



(a) `AALEELVK` peptide from the `P12081ups|SYHC_HUMAN_UPS` protein.

(b) `VLPLIIPILSK` peptide from the `sp|F4I893|ILA_ARATH` protein.

Figure 5.3: Posterior distributions of the difference of means between the 0.05 fmol UPS spike condition (μ_1) and the 10 fmol UPS spike condition (μ_7) and the corresponding boxplots summarising the observed data. The 95% credible interval is indicated by the blue central region.

(Figure 5.3b) suggests that the probability that means differ is low. Those conclusions support the summaries of raw data depicted on the bottom panel of Figure 5.3. Moreover, the posterior distribution provides additional insights on whether a peptide is under-expressed or over-expressed in a condition compared to another. For example, looking back to the UPS peptide, Figure 5.3a suggests an over-expression of the `AALEELVK` peptide in the seventh group (being the condition with the highest amount of UPS spike) compared to the first group (being the condition with the lowest amount of UPS spike), which is consistent with the experimental design. Furthermore, the middle panel merely highlights the fact that the posterior distribution of the difference $\mu_1 - \mu_7$ is the symmetric of $\mu_7 - \mu_1$, thus the sense of the comparison only remains an aesthetic choice.

5.4.2 The benefit of intra-protein correlation

One of the main benefits of our methodology is to account for between-peptides correlation, as described in Section 5.2. As the first illustration of such property, we modelled

correlations between all quantified peptides derived from the same protein. In order to highlight the gains that we may expect from such modelling, we displayed on Figure 5.4 the comparison between a differential analysis using our univariate method or using the multivariate approach. Recall the quantification data from the previous subsection. In this

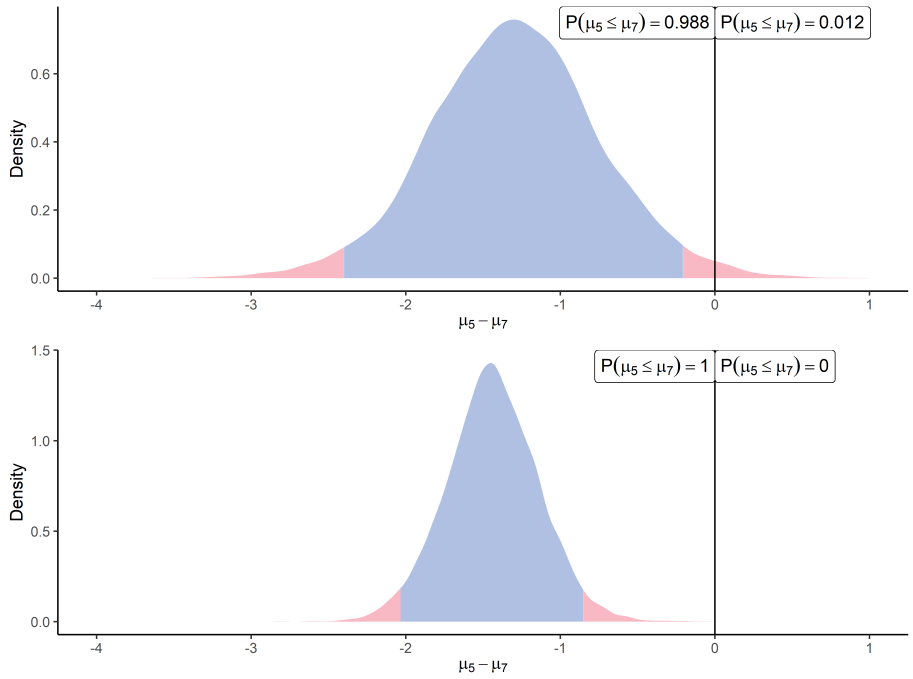


Figure 5.4: Posterior distributions of the mean difference $\mu_5 - \mu_7$ for the `AALEELVK` peptide from the `P12081ups|SYHC_HUMAN_UPS` protein using the univariate approach (top) and the multivariate approach (bottom). The 95% credible interval is indicated by the blue central region.

example, we purposefully considered a group of 9 peptides coming from the same protein (`P12081ups|SYHC_HUMAN_UPS`), which intensities may undoubtedly be correlated to some degree. We consider in this section the comparison of intensity means between the fifth point (2.5 fmol UPS - $\boldsymbol{\mu}_5$) and the seventh point (10 fmol UPS - $\boldsymbol{\mu}_7$) of the UPS spike range. The posterior difference of the mean vector $\boldsymbol{\mu}_5 - \boldsymbol{\mu}_7$ between two conditions has been computed, and the first peptide (`AALEELVK`) has been extracted for graphical visualisation. Meanwhile, the univariate algorithm has also been applied to compute the posterior difference $\mu_5 - \mu_7$, solely on the peptide `AALEELVK`. The top panel of Figure 5.4 displays the latter approach, while the multivariate case is exhibited on the bottom panel. One should observe clearly that, while the location parameter of the two distributions is close as expected, the multivariate approach takes advantage of the information coming from the correlated peptides to reduce the uncertainty in the posterior estimation. This lower variance provides a tighter range of probable values, enabling a more precise estimation of the effect size and increased

confidence in the resulting inference (deciding whether the peptide is differential or not).

5.4.3 The mirage of imputed data

After discussing the advantages and the valuable interpretative properties of our methods, let us mention a pitfall that one should avoid for the inferences to remain valid. In the case of univariate analysis, we pointed out thanks to Equation (5.4) that all the useful information is contained on observed data, and no imputation is needed since we already integrated out all missing data. Imputation does actually not even make sense in one dimension since, by definition, a missing data point is simply equivalent to an unobserved one, and we shall gain more information only by collecting more data. Therefore, one should be really careful when dealing with imputed datasets and keep in mind that imputation somehow *creates* new data points that do not bear any additional information. Thus, there is a risk of artificially decreasing the uncertainty of our estimated posterior distributions simply by considering more data points in the computations than what was genuinely observed. For

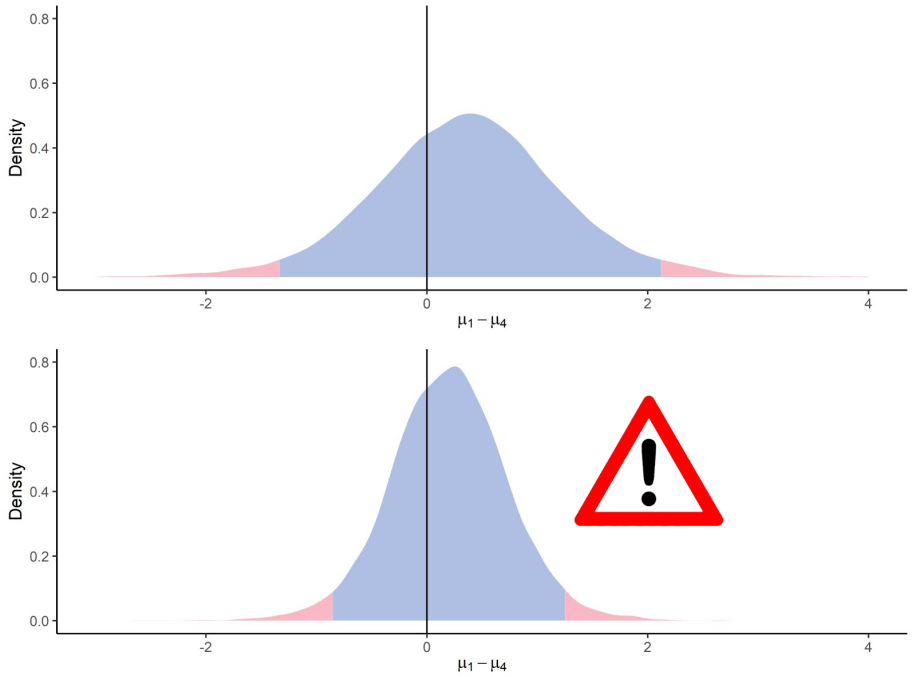


Figure 5.5: Posterior distributions of the mean difference $\mu_1 - \mu_4$ for the `EVQELAQEAAER` peptide from the `sp|F4I893|ILA_ARATH` protein using the observed dataset (top) and the imputed dataset (bottom). The 95% credible interval is indicated by the blue central region.

instance, imagine a dummy example where 10 points are effectively observed, and 1000 remain missing. It would be a massive error and underestimation of the true variance to impute the 1000 missing points (say with the average of the ten observed ones) and use

the resulting 1010-dimensional vector for computing the posterior distributions of the mean. Let us mention that such a problem is not specific to our framework and more generally also applies to Rubin's rules. One should keep in mind that those approximations only holds for a reasonable ratio of missing data. Otherwise, one may consider adapting the method, for example, by penalising the degree of freedom in the relevant t-distributions. To illustrate this issue, we displayed on Figure 5.5 an example of our univariate algorithm applied both on the observed dataset (top panel) and the imputed dataset (bottom panel). In this context, we observe a reduced variance for the imputed data. However, this behaviour is just an artefact of the phenomenon mentioned above: the bottom graph is merely not valid, and only raw data should be used in our univariate algorithm to avoid spurious inference results. More generally, while imputation is sometimes needed for the methods to work, one should always keep in mind that it always constitutes a bias (although controlled) that should be accounted for with tailored solutions, as this manuscript intends to provide.

5.4.4 Acknowledging the effect size

After discussing methodological aspects, let us dive into more biological-related properties displayed on Figure 5.6. The three panels describe the increasing differences that can be

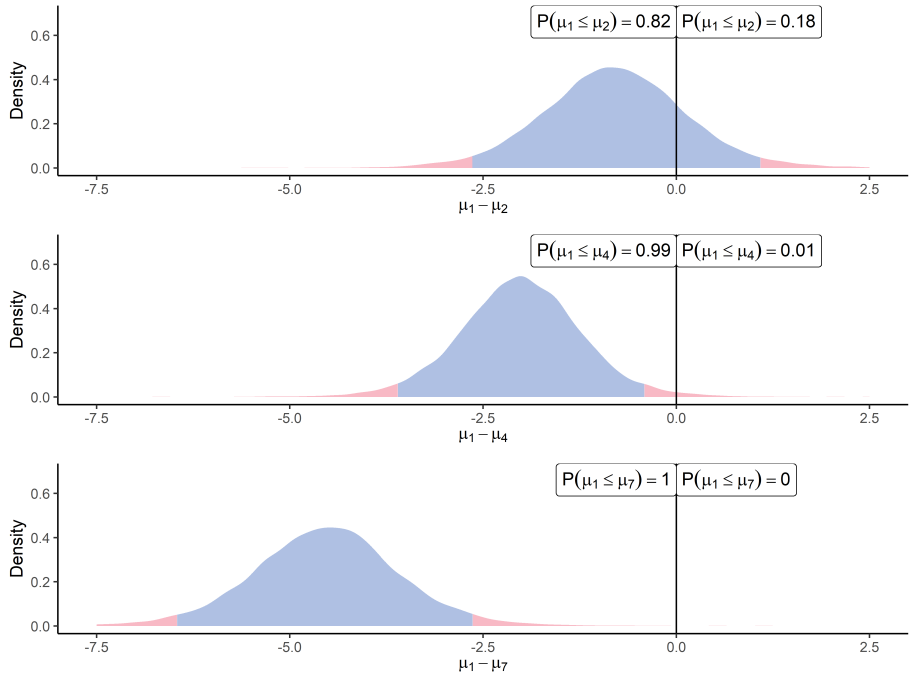


Figure 5.6: Posterior distributions of the mean differences $\mu_1 - \mu_2$, $\mu_1 - \mu_4$ and $\mu_1 - \mu_7$ for the `AALEELVK` peptide from the `P12081ups|SYHC_HUMAN_UPS` protein. The 95% credible interval is indicated by the blue central region.

observed when we compare sequentially the first point (0.05 fmol UPS) of the UPS spike range (μ_1) to the second one (0.25 fmol UPS - μ_2), the fourth one (1.25 fmol UPS - μ_4) and the highest one (25 fmol UPS - μ_7). The experimental design suggests that the difference in means for a UPS peptide should increase with respect to the amount of UPS proteins that was spiked in the biological sample (see Chapter 2). This illustration offers a perspective on how this difference becomes more and more noticeable, though mitigated by the inherent variability. Such an explicit and adequately quantified variance, and the induced uncertainty in the estimation, should help practitioners to make more educated decisions with the appropriate degree of caution. In particular, Figure 5.6 highlights the importance to consider the effect size (increasing here), which is crucial when studying the underlying biological phenomenon. Such a graph may recall us that statistical inference should be more about offering helpful insights to experts of a particular domain, rather than defining automatic and blind decision-making procedures (Betensky, 2019). Moreover, let us point out that current statistical tests used for differential analysis express their results solely as p-values. One should keep in mind that, no matter their value, they do not provide any information about the effect size of the phenomenon (Sullivan and Feinn, 2012).

5.4.5 About protein inference

To conclude on the practical usage of the proposed multivariate algorithm, let us develop ideas for comparing simultaneously multiple peptides or proteins. As highlighted before, accounting for the covariances between peptides tends to reduce the uncertainty on the posterior distribution of a unique peptide. However, we only exhibited examples comparing one peptide at a time between two conditions, although in applications, practitioners often need to compare thousands of them simultaneously. From a practical point of view, while possible in theory, we probably want to avoid modelling the correlations between every combination of peptides into a full rank matrix for at least two reasons.

First, it probably does not bear much sense to assume that all peptides in a biological sample interact with no particular structure. Secondly, it appears unreasonable to do so from a statistical and practical point of view. Computing and storing a matrix with roughly 10^4 rows and columns induces a computational and memory burden that would complicate the procedure while potentially leading to unreliable objects if matrices are estimated merely on a few data points, as for our example. However, a more promising approach would consist in deriving a sparse approach by levering the underlying structure of data from a biological perspective. If we reasonably assume, as before, that only peptides from common proteins present non-negligible correlations, it is then straightforward to define a block-diagonal matrix for the complete vector of peptides, which would be far more reasonable to estimate. Such an approach would take advantage of both of our algorithms by using the factorisation (as in Equation (5.4)) over thousands of proteins to sequentially estimate a high number of low dimensional mean vectors. Assuming an example with a thousand proteins

containing ten peptides each, the approximate computing and storage requirements would be reduced from a $(10^4)^2 = 10^8$ order of magnitude (due to one high-dimensional matrix) to $10^3 \times 10^2 = 10^5$ (a thousand of small matrices). In our applicative context, the strategy of dividing a big problem into independent smaller ones appear beneficial from both the applicative and statistical perspective.

This being said, the question of the *global* inference, in contrast with a peptide-by-peptide approach, remains pregnant. To illustrate this topic, let us provide on Figure 5.7 an example of simultaneous differential analysis for nine peptides from the same protein. According to our previous recommendations, we accounted for the correlations through the multivariate algorithm and displayed the results in posterior mean's differences for each peptide from the `P12081ups|SYHC_HUMAN_UPS` protein at once (*i.e.* $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_7$). In this example, eight peptides over nine contained in the protein are clearly differential in the same direction with comparable effect sizes, corroborating our intuition of correlated quantities. However, the situation may become far trickier when distributions lie closer to 0 on the x-axis or if only one peptide presents a clear differential pattern. As multiple and heterogeneous situations could be encountered, we do not provide here recommendations for directly dealing with protein-scale inference. Once again, the criterium for deciding what should be considered as *different enough* is highly dependent on the context and reasonable hypotheses, and no arbitrary threshold may bear any kind of general relevancy. However, we should still point out that our Bayesian framework provides convenient and natural interpretations in terms of a probability for each peptide individually. It is then straightforward to construct probabilistic decision rules and combine them to reach a multivariate inference tool, for instance, by computing an average probability for the means' difference to be below 0 across all peptides. However, one should note that probability rules prevent directly deriving global probabilistic statements without closely looking at dependencies between the single events (for instance, the factorisation in Equation (5.4) holds thanks to the induced independence between peptides). Although such an automatic procedure cannot replace the expert analysis, it may still provide a handy tool for extracting the most noteworthy results from a massive number of comparisons, which the practitioner should look at more closely afterwards. Therefore, once a maximal risk of the adverse event or a minimum probability of the desired outcome has been defined, one may derive the adequate procedure to reach those properties.

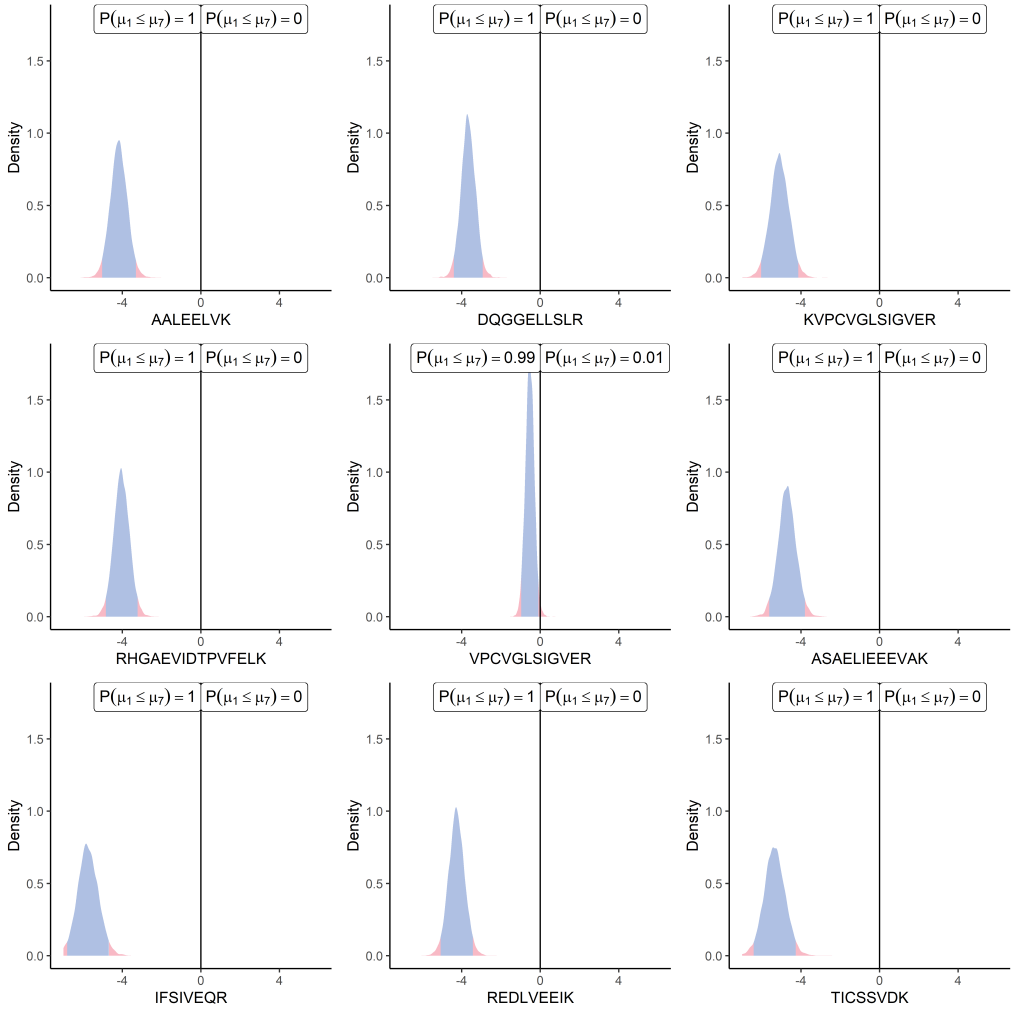


Figure 5.7: Posterior distributions of mean difference $\mu_1 - \mu_7$ for the nine peptides from the `P12081ups|SYHC_HUMAN_UPS` protein using the multivariate approach. The 95% credible interval is indicated by the blue central region.

5.5 Conclusion and perspectives

This chapter presents a Bayesian inference framework to tackle the problem of differential analysis in both univariate and multivariate context, while accounting for possible missing data. We proposed two algorithms, levering classical results from conjugate priors to compute posterior distributions and easily sample the difference of means when comparing groups of interest. For handling the recurrent problem of missing data, our multivariate approach takes advantage of the multiple imputations' approximation, while the univariate framework allows us to simply ignore this issue. In addition, this methodology aims at providing information not only on the probability of the means' difference to be null, but also on the uncertainty quantification as well as the effect sizes, which are crucial in a biological framework.

We believe that such probabilistic statements offer valuable inference tools to the practitioners. In the particular context of differential proteomics, this methodology allows us to account for between-peptides correlations. With an adequate decision rule and an appropriate correlation structure, Bayesian inference could be used in large-scale proteomics experiments, such as label-free global quantification strategies. Nevertheless, targeted proteomics experiments could already benefit from this approach, as the set of considered peptides is restricted. Furthermore, such experiments used in biomarker research could greatly benefit from the quantification of the uncertainty and the assessment of the effect sizes.

Although promising and illustrated on real applicative problems, this work still remains under development and would necessitate a further extensive simulation study for assessing more precisely the properties of the method. Readers could also benefit from more insights about practical usage, by providing intuitions for calibration of the hyper-parameters or precise estimations of the expected running times. Finally, while we considered the influences at a protein-scale, introducing correlations according to different biological features would represent an interesting path to explore.