A Methodology for Two-Level Product Partition Model Estimation of Normal Means

Paul Gerard Diver
Washington, DC

M.A., Economics, University of Virginia, 2010
M.S., Mathematics and Statistics, Georgetown University, 2007
B.S., Mathematics, Georgetown University, 2006

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Statistics

University of Virginia
May, 2017

A Methodology for Two-Level Product Partition Model Estimation of Normal Means

Dissertation

Paul Diver

May, 2017

University of Virginia

Department of Statistics

Committee

Professor Jeff Holt

Professor Karen Kafadar

Professor Gretchen Martinet

Professor Nolan Wages

Contents

Abstract

In many instances, a collection of items can be thought of as having two levels: an *individual-level* in which each item is uniquely identified and a *group-level* defined by a set of known group membership labels. A two-level mean estimation and clustering problem using probability models is addressed where each item has an independent observation following a normal distribution with an item specific unknown mean and constant variance. The two-level structure allows the individual-level item observations to be aggregated and averaged by group membership index. This implies that each group index has an associated mean equal to the average of its members' means. This two-level setting is studied adapting probability models which allow means to be equal at both the individual and group-levels. The possibility of equal means at the group-level implies a *two-level mean condition* which restricts the possible values of the means to be estimated and thus is necessarily incorporated into the model. These probability models, called product partition models, provide a logical and flexible framework to the problem and permit the use of popular computational tools. Given a set of items, a partition is an arrangement of these items into a collection of non-empty, non-overlapping subsets. Items in the same set have observations with the same mean. Similarly, the group indices may also be partitioned into group-level sets. Random partitions at the group and individual-levels jointly possess a probability distribution which is updated through the information contained in the data. This posterior distribution allows for the estimation of the unknown means. Markov sampling adapted for the two-level setting assists in providing two estimates: a *two-level product estimate* computed via a weighted average of posterior means summed over all possible pairs of group and individual-level partitions, and an estimate via the posterior mode, the maximum *a posteriori* clustering of the data. The posterior mode

provides a clustering structure at both the individual and group-levels, herein automatically selecting the number of clusters at both levels and avoiding the need for presetting as with other methods. This dissertation extends the insightful work of Crowley (1997) to this two-level setting with the two-level mean condition. Her important work focuses on estimating the means of normally distributed observations with known variance equal to 1 in the one-level setting where no group-level index information is utilized, whether or not it is known. The incorporation of this two-level mean condition provides not only logically consistent analysis regarding the means across levels, but also superior mean estimation results including when applied to Major League Baseball batting average data studied in Crowley (1997) and elsewhere.

1 Introduction

In many instances, a collection of items can be thought of as having two levels: an *individual-level* in which each item is uniquely identified and a *group-level* defined by a set of known group membership labels. Each individual item possesses a known group membership. Each item may be uniquely identified by a double-index, ij where i indexes group membership and j provides an index of individuals within a group. An example of this structure can be seen in Congressional representation where the individual representatives constitute the individual-level items and the states they represent are the group labels.

The present document addresses a two-level mean estimation and clustering problem using probability models where each item (ij) has an associated independent observation $\mathrm{Y}_{ij} \sim N(\mu_{ij}, \sigma^2)$. The two-level structure allows the individual-level item observations to be aggregated and averaged by group membership index. These group-level statistics each have a normal distribution and an implied mean related to the distribution means corresponding to the individuals in each group. To wit, each group label i has an associated implied mean $\mu_i = n_i^{-1} \sum_{j=1}^{n_i} \mu_{ij}$ where n_i is the number of items with group label i. These implied means constitute group-level means.

This two-level setting is studied using probability models which allow means to be equal at both the individual and group levels. The possibility of equal means at the

group-level implies a *two-level mean condition* such that if $\mu_i = \mu_{i'}$ then

$$\frac{1}{n_i}\sum_{j=1}^{n_i}\mu_{ij} = \mu_i = \mu_{i'} = \frac{1}{n_{i'}}\sum_{j'=1}^{n'_i}\mu_{i'j'} \tag{1.1}$$

This two-level mean condition restricts the possible values of the means to be estimated and thus is necessarily incorporated into the model.

Product partition models (PPMs) provide a logical framework to the problem, facilitate the explanation of the setting, and yield conveniently tractable solutions for many scenarios. Moreover, popular computational tools may be employed as a result of the design. Additionally, when a clustering structure is desired, it may be inferred from the analysis. Unlike other popular clustering methods, no preselection of an exact or a maximum number of clusters is necessary.

Product partition models are frequently used to estimate parameters and identify unknown structures of interest in a variety of data settings. As their name implies, these are product-form models which provide a convenient partition-based analytical framework in which to work. Given a set of items or objects, a partition is an arrangement of these items into a collection of non-empty, non-overlapping subsets. Random partitions possess a probability distribution, as do the observations (the data) corresponding to the items.

At the individual-level, the individual items may be partitioned into sets. Items in the same set have observations which follow normal distributions with the same mean. Accordingly, these partition sets may be viewed as clusters of similar items. Similarly, the group indices may also be partitioned into group-level sets where μ_i's corresponding to indices in the same set are equal.

When considering partitions at *both* the individual and group-levels, failing to appropriately account for the implied relationships between the group and individual-

level means across these levels can lead to logically inconsistent mean estimation. As noted, if group indices i and i' are clustered into the same group-level partition set, then $\mu_i = \mu_{i'}$. This implies the *two-level mean condition* defined above in (1.1).

Given group and individual-level partitions of the items, denoted by ρ_{g} and ρ_{ind}, respectively, this two-level mean condition restricts the possible values of the μ_{ij}'s. Indeed, certain means may necessarily be linear combinations of others. Each pair of group and individual-level partitions has a corresponding set of restrictions on the means.

This dissertation extends the insightful work of Crowley (1997) [Cro97] to this two-level setting with the two-level mean condition. Her important work focuses on estimating the means of normally distributed observations with known variance equal to 1 in the *one-level* setting where no group-level index information is utilized, whether or not it is known.

The present exposition concerns estimating the μ_{ij}'s and simultaneously identifying suitable partitions ρ_{g} and ρ_{ind} of the items at the group and individual-levels, respectively, while requiring the two-level mean condition to hold throughout the analysis. This is accomplished by adapting product partition models to the two-level setting and identifying how to incorporate the implications of the two-level mean condition into the model.

This document provides two estimates of the means: a *two-level product estimate* and an estimate of the posterior mode. The two-level product estimate is a weighted average of posterior means conditioned on ρ_{g}, ρ_{ind}, and the data in a manner which ensures the two-level mean condition is satisfied. This extends the concept of a one-level product estimate as used in [Cro97] to the two-level setting. The posterior mode is the set of posterior means associated with the maximum *a posteriori* pair of partitions. Accordingly, a specific clustering of the items at both the group and

individual-level is implied with the posterior mode estimate.

Background

Introduced by Hartigan (1990) [Har90], PPMs have been used for clustering and inferential purposes in a wide variety of theoretical and applied contexts. These probability models allow data to weight probable partitions and subsequently draw inference for prediction. Additionally, this paper demonstrates the capacity and flexibility of PPMs in metadata analysis. Barry and Hartigan (1992) [BH92] use PPMs in a variety of change point problems on sequential observations. Together, these documents constitute the acknowledged foundation of the PPM literature. A wide variety of extensions followed and provide a robust field of work. Notably, Crowley (1992) [Cro92] and [Cro97] address a univariate normal means estimation problem with known variance. Each of these papers, as do the earlier Hartigan works, define and make use of a product estimate term (one-level) which provides a weighted average estimate of item observation means across all possible partitions of associated items. A suite of works, Loschi et al. (1999) [LIAV99], Loschi, Cruz, and Arellano-Valle (2005) [LCAV05], and Loschi, Moura, and Iglesias (2005) [LMI05], extend results addressing normal means to product estimates concerning normal variances and further more generally to parameters of the regular exponential family. This type of mean estimate is used in varying capacities in subsequent literature and is featured heavily in later chapters in this document.

Quintana and Iglesias (2003)[QI03] introduces a decision theoretic approach to model selection and explores outlier detection in regression models. This influential paper is perhaps most notable however for showing a connection between Dirichlet processes and PPMs through the prior distribution on the partition. These results formally connect and provide an explicit basis for the prior partition distribution

used by [Cro97] and many other entries in the PPM literature. This prior, recently called the Ewens-Pitman prior, as well as several other popular PPM priors are discussed and compared in Casella, Morena, and Giron (2014) [CMG$^+$14] emphasizing the importance of appropriate prior selection. Prior distribution selection in the two-level setting is examined later in this dissertation, notably comparing different combinations of partition priors as multiple priors are allowed in the two-level model as opposed to one in the one-level setting.

A number of papers utilize PPMs to address classification problems utilizing predictor variables including Holmes et al. (1999) [HDM99], Denison and Holmes (2001) [DH01], and Denison et al (2002) [DAHH02]. Predictors as random variables are presented in Holmes et al. (2005) [HDRM05] and in a Bayesian generalized PPM provided by Park and Dunson (2010) [PD10] through their insertion into nonnegative functions called *cohesions*. These functions are used throughout the PPM literature and are vital to the construction of many prior partition distributions. Jordan, Livingstone, and Barry (2007) [JLB07] use PPMs to estimate success probabilities (of a binary response variable) from patterns of predictor variables. Dahl (2009) [D$^+$09] provides a novel algorithm to quickly find the maximum *a posteriori* (MAP) and maximum likelihood (ML) clustering solutions for a specific class of univariate product partition models with ordered data. Further work in PPMs with covariates in a variety of settings including as predictors in augmented response vectors and in the incorporation of a covariate *similarity* function into cohesions is developed in Muller et al. (2008) [MQR08] and Park and Dunson (2010) [PD10]. Muller and Quintana (2010) [MQ10] and Muller, Quintana, and Rosner (2011) [MQR11]suggest what they term a PPMx model which also incorporates a similarity function directly into the cohesions. These papers discuss models in which the *a priori* partition probabilities may be affected through covariate information. Continuing this focus more

recently are Costa el al. (2013) [CGB+13] and Dahl, Day, and Tsai (2016) [DDT16] with the notable addition that they address models when certain exchangeability assumptions in the items are not enforced or do not hold. It is worth mentioning that throughout the PPM literature many of the authors laud the straightforward nature of PPMs for their utility in addressing such a wide variety of settings.

Dissertation Organization

Chapter 2 presents the data and the two-level layout. This chapter describes the two-level partition structure and introduces how the implied mean relationships resulting from the partitions and the two-level mean condition relate to the μ_{ij}'s. Chapter 3 describes the two-level model. Through a matrix $\mathbf{H}$ which will be defined later in this dissertation, one can track the influence of the two-level mean condition throughout the model. Herein two observation settings are presented: $\mathcal{A}$ the observations have constant and known variance, and $\mathcal{B}$ the observations have constant and unknown variance. Chapter 4 presents how to encode the restrictions implied by the two-level mean condition through the mentioned $\mathbf{H}$ matrix. Chapter 5 presents the computational approach to estimating the observation means. The two estimates of the μ_{ij}'s, the two-level product estimate and the posterior mode, are discussed in this chapter. As mentioned above, the two-level product estimate is computed over all possible partition pairs of ρ_{g} and ρ_{ind}. For all but a very small number of items and groups, it becomes impractical to compute this term exactly due to the quickly increasing number of possible partitions. As such, a Markov sampling scheme, adapted for the two-level setting, is used for computing these estimates and is detailed here. In Chapter 6, the model is demonstrated on simulated data. This chapter discusses how to address appropriately the hyperparameters noted in Chapter 3. Superior results over the one-level method presented in [Cro97] and

the popular mixture-model method of Fraley and Raftery (2002) [FR02] and Fraley, Raftery, and Scrucca (2012) [FRS12] are illustrated. Additionally, the robustness of the model to hyperparameter specification and model assumptions, namely distribution misspecification and the constant variance assumption, are discussed. As stated above and described in [CMG+14], in the one-level setting, results have been shown to be quite sensitive to the selection of the partition prior distribution. Chapter 7 provides guidance for the two-level setting considering combinations of classes of partition priors at the group and individual-levels. Chapter 8 applies the two-level PPM and associated product estimate to the Major League Baseball batting data used in [Cro97] and George (1986) [Geo86]. The two-level product estimate provides a lower average squared loss than [Cro97] and many other estimates of future team batting averages. Chapter 9 discusses directions for future research including an example for extending the two-level PPM with the two-level mean condition to the multivariate setting.

2 Relevant Data and the Two-Level Layout

Introduction

When a collection of individual items has a known group structure, a *two-level* analysis is possible as opposed to a one-level analysis based solely on individual-level information. Incorporating the group-level information may imply conditions which restrict parameter values being estimated. Failing to model these conditions appropriately can lead to inferiorly specified models and logically inconsistent results. Conversely, the careful construction of a model incorporating these conditions can lead to a better understanding of the data and more meaningful results.

Notation

Let there be a collection of items, Ω_1. Each item is uniquely identified by a double-index, ij. The first index, i, provides an indicator of a known group membership (for example, state or team). The second index, j, provides an index of individuals within a group (for example, the j^{th} elected official of state i or the j^{th} player on team i). Let there be g groups with $n_i,\ i = 1, ..., g$, individuals in each respective

group. Therefore, Ω_1, the collection of items, may be defined as:

$$(ij) \in \Omega_1 = \{(i,j) : \ i = 1, ..., g; \ j = 1, ..., n_i\}$$

Let each unique item (ij) have an associated observation, Y_{ij}. As in [Cro97],[1] let the observations be independent and follow a Normal distribution such that Y_{ij} has mean μ_{ij} and variance,[2] σ^2:

$$Y_{ij}|\mu_{ij} \sim N(\mu_{ij}, \sigma^2), \qquad (ij) \in \Omega_1$$

Let $\mathbf{Y} = \left[Y_{11}, Y_{12}, ..., Y_{1n_1}, Y_{21}, ..., Y_{gn_g}\right]^T$ and define $\boldsymbol{\mu}$ to be a vector such that

$$\boldsymbol{\mu} = \left[\mu_{11}, \mu_{12}, ..., \mu_{1n_1}, \mu_{21}, ..., \mu_{gn_g}\right]^T$$

Let an *individual-level partition*, ρ_{ind}, be a mapping of each item (ij) to some set $S_s, \ s = 1, ..., k$, such that $\rho_{\text{ind}} = \{S_1, S_2, ..., S_k\}$ with $S_s \cap S_{s'} = \emptyset, \ \forall s \neq s'$, and $\cup_s S_s = \Omega_1$. A mapping function k_1 is defined such that if item $(ij) \in S_s$, then $k_1(ij) = S_s$. This function is a clustering function in that if $k_1(ij) = k_1(i'j') = S_s$ for $(ij) \neq (i'j')$, then $\mu_{ij} = \mu_{i'j'} = \mu^{S_s}$. The sets $S_s, \ s = 1, ...k$, can be equivalently thought of and referred to as clusters of items. Define $\boldsymbol{\mu}_{\text{var}}$ such that it is the vector of the individual-level set means, $\boldsymbol{\mu}_{\text{var}} = \left[\mu^{S_1}, \mu^{S_2}, ..., \mu^{S_k}\right]^T$. Accordingly, each $\mu_{ij} \in \boldsymbol{\mu}_{\text{var}}$.

Through the known group indices, the items may be aggregated and considered at the group-level. Let Ω_2 be the collection of group-level indices, $i \in \Omega_2 = \{i : i = 1, ..., g\}$. The individual-level item observations may be averaged by these group

[1] To be clear, [Cro97] does not incorporate group-level information and performs a one-level analysis. The items, observations, and corresponding means have a single index which uniquely identifies each component.

[2] [Cro97] assumes a known and constant $\sigma^2 = 1$.

indices defining group-level statistics for each index i

$$Y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

It follows that these group-level statistics are normally distributed with means implied by the distribution means corresponding to the individuals in each group. That is to say, each group index i has an associated implied mean

$$\mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mu_{ij}$$

These implied means constitute group-level means. Accordingly, the group indices may also be partitioned into group-level sets when these group-level means are in common. Each group-level index, $i \in 1, ..., g$, may be mapped to a set $T_t,\ t = 1, ..., l$, forming a *group-level partition*, ρ_{g}, defined such that $\rho_{\mathrm{g}} = \{T_1, T_2, ..., T_l\}$ with $T_t \cap T_{t'} = \emptyset,\ \forall t \neq t'$, and $\cup_t T_t = \Omega_2$. Similar to the individual-level, a mapping function k_2 is defined such that if $i \in T_t$, then $k_2(i) = T_t$. If $k_2(i) = k_2(i') = T_t$ for $(i) \neq (i')$, then $\mu_i = \mu_{i'} = \mu^{T_t}$. These group-level partition sets $T_t,\ t = 1, ..., l$, may be thought of as clusters of group-level indices.

The group indices provide a *two-level* structure to the data which has implications regarding the μ_{ij}'s. If $k_2(i) = k_2(i') = T_t$, then not only does $\mu_i = \mu_{i'} = \mu^{T_t}$, but the *two-level mean condition*, an integral part of the new model, implies that

$$\frac{1}{n_i} \sum_{j=1}^{n_i} \mu_{ij} = \mu_i = \mu^{T_t} = \mu_{i'} = \frac{1}{n_{i'}} \sum_{j'=1}^{n'_i} \mu_{i'j'} \tag{2.1}$$

As a result of the two-level mean condition, given ρ_{g} and ρ_{ind}, the group indices imply additional linear relationships between the individual-level set means, and thus restrict the possible values of these means. Other than these restrictions implied

by the two-level mean condition, there are no other restrictions on the set means.

Accordingly, $\boldsymbol{\mu}_{\text{var}}$ may be broken down into complementary subsets of individual-level set means, $\boldsymbol{\mu}_{\text{rel}}$ and $\boldsymbol{\mu}_0$, such that each element of $\boldsymbol{\mu}_{\text{rel}}$ is a linear combination of the elements of $\boldsymbol{\mu}_0$ with $\boldsymbol{\mu}_0 \cup \boldsymbol{\mu}_{\text{rel}} = \boldsymbol{\mu}_{\text{var}}$. It is shown in Chapter 4 that the vector of individual-level item observation means, $\boldsymbol{\mu}$, may be defined in terms of the elements of $\boldsymbol{\mu}_0$, the free individual-level set means, through a mapping matrix $\mathbf{H}$ such that $\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\mu}_0$. This matrix $\mathbf{H}$ encodes all of the mean relationships resulting from ρ_{g}, ρ_{ind}, and the two-level mean condition.

The present exposition concerns estimating the μ_{ij}'s and simultaneously identifying suitable partitions, ρ_{g} and ρ_{ind}, given the data while requiring the two-level mean condition to hold throughout the analysis. This is done by adapting product partition models to the two-level setting and identifying how to incorporate the implications of the two-level mean condition into the model. Building the model in this manner yields not only a more insightful analysis of the data than failing to do so, but also extends the logical consistency of the two-level mean condition to the estimates of the means given any pair of partitions ρ_{g} and ρ_{ind}. This can be seen explicitly in the posterior mode estimates of the μ_{ij}'s.

Incorporating a condition on the variance terms was considered as well, but is avoided for several reasons. Note that when individual-level observation variances are constant, $n_i^{-1} \sum_{j=1}^{n_i} Y_{ij} = Y_i \sim N(\mu_i, \sigma^2/n_i)$. If all the group sizes are equal, $n_i = n \; \forall i$, then any condition requiring the variances corresponding to group indices in the same group-level set to be equal would be automatically satisfied. If the group sizes are unbalanced, then n_i would have to equal $n_{i'}$ in order for $k_2(i) = k_2(i') = T_t$. The group-level partition structure would be overly sensitive to the group sizes. In turn, the individual-level mean relationships would be affected through the two-level mean condition. In the event that all groups were of different sizes, all group in-

dices would be mapped to different sets, and the two-level mean condition would not provide any additional information beyond that of the one-level situation even if multiple group-level means were the same. As such, any explicit variance condition is unhelpful.

3 The Model

The joint distribution of the group-level partition, the individual-level partition, and the data can be decomposed into

$$P(\rho_{\mathrm{g}},\ \rho_{\mathrm{ind}},\ \mathbf{Y}) = P(\rho_{\mathrm{g}}, \rho_{\mathrm{ind}}) \times p(\mathbf{Y}|\rho_{\mathrm{g}},\ \rho_{\mathrm{ind}}) \tag{3.1}$$

where $P(\rho_{\mathrm{g}}, \rho_{\mathrm{ind}})$ is the joint prior distribution of the group and individual-level partitions, and $p(\mathbf{Y}|\rho_{\mathrm{g}},\ \rho_{\mathrm{ind}})$ is the conditional density of $\mathbf{Y}$ given the partitions ρ_{g} and ρ_{ind}.

The joint prior distribution on the partitions, $P(\rho_{\mathrm{g}}, \rho_{\mathrm{ind}})$, may be addressed by specifying the marginal prior partition distributions and using their independence to derive their joint distribution, $P(\rho_{\mathrm{g}}, \rho_{\mathrm{ind}}) = P(\rho_{\mathrm{g}})P(\rho_{\mathrm{ind}})$. In general, there are a number of ways that the marginal and joint prior distributions on the partitions could be specified. In the context of a product partition model (PPM) framework however, as the name implies, partitions have associated product models ([Har90]). Accordingly, define the prior distributions for the group and individual-level partitions respectively to be:

$$P(\rho_{\mathrm{g}} = \{T_1, T_2, ..., T_l\}) = L \prod_{t=1}^{l} c_T(T_t) \tag{3.2}$$

$$P(\rho_{\mathrm{ind}} = \{S_1, S_2, ..., S_k\}) = K \prod_{s=1}^{k} c_S(S_s) \tag{3.3}$$

where c_T and c_S are group and individual-level *cohesion* functions such that $c_T(T)$

and $c_S(S) \geq 0,\ \forall T, S.$ L and K are normalization constants such that the probabilities of all possible partitions at their respective levels sum to 1. The Ewens-Pitman partition prior is most common in the PPM literature, and indeed a variation of it is used in [Cro97]. A proof of the equivalency of the initial prior partition distribution used in [Cro97] and the more standard Ewens-Pitman form is provided at the end of this chapter (Proof 3.1). For such a prior,

$$c(S_s) = m * (n_{S_s} - 1)!$$

where m is a hyperparameter which is set or estimated, and n_{S_s} is the number of items in set S_s of the partition.[1] This prior is given the greatest attention here, though several other notable priors, the Uniform, Constant, and Hierarchical Uniform prior distributions, are considered as well.

The conditional density of $\mathbf{Y}$ given the partitions ρ_{g} and ρ_{ind}, $p(\mathbf{Y}|\rho_{\text{g}},\ \rho_{\text{ind}})$, is found by integrating out the parameters $\boldsymbol{\phi}$ given in the data distribution and informed by a prior:

$$p(\mathbf{Y}|\rho_{\text{g}}, \rho_{\text{ind}}) = \int p(\mathbf{Y}|\boldsymbol{\phi},\ \rho_{\text{g}},\ \rho_{\text{ind}}) f(\boldsymbol{\phi}|\rho_{\text{g}},\ \rho_{\text{ind}}) d\boldsymbol{\phi} \tag{3.4}$$

As stated earlier, each Y_{ij} follows a Normal distribution with mean μ_{ij} and variance σ^2, and the observations are independent. Accordingly,

$$p(\mathbf{Y}|\boldsymbol{\mu}, \sigma^2,\ \rho_{\text{g}},\ \rho_{\text{ind}}) \sim MVN(\boldsymbol{\mu}, \boldsymbol{I}^*) \tag{3.5}$$

where $\boldsymbol{I}^* = \sigma^2 \boldsymbol{I}_{q \times q}$ and $q = \sum_{i=1}^{g} n_i$. As noted above, $\boldsymbol{\mu}$ may be defined in terms of a subset of the individual-level set means, $\boldsymbol{\mu}_0$, by $\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\mu}_0$ thus encoding the linear

[1] The use of the notation S_s should be thought of as general here. To be clear, this cohesion form applies similarly to group-level partitions where sets are noted T_t.

relationships between the individual-level set means implied by the given partitions and the two-level mean condition. As such, (3.5) may be restated in terms of $\boldsymbol{\mu}_0$:

$$p(\mathbf{Y}|\boldsymbol{\mu}_0, \sigma^2,\ \rho_{\text{g}},\ \rho_{\text{ind}}) \sim MVN(\mathbf{H}\boldsymbol{\mu}_0, \boldsymbol{I}^*) \tag{3.6}$$

By specifying a conjugate prior distribution for f, $p(\mathbf{Y}|\rho_{\text{g}}, \rho_{\text{ind}})$ may be found in closed form. (3.4) is computed for two specific settings below:

Setting $\mathcal{A}$: Known and Constant Observation Variance

When σ^2 is known, let the prior distribution on the free mean vector be such that $\boldsymbol{\mu}_0 \sim MVN(\boldsymbol{\theta}^*, \boldsymbol{\gamma_0})$, where $\boldsymbol{\theta}^*$ and $\boldsymbol{\gamma_0}$ are estimated from the data[2] or specified according to the preferences of the modeler. The covariance matrix $\boldsymbol{\gamma_0}$ is of the form $\text{c}\mathbf{I}_w$ where c is a constant, and $w = ||\boldsymbol{\mu}_0||$, the cardinality of $\boldsymbol{\mu}_0$. Suppressing σ^2 in the notation, (3.4) becomes

$$p(\mathbf{Y}|\rho_{\text{g}}, \rho_{\text{ind}}) = \int p(\mathbf{Y}|\boldsymbol{\mu}_0,\ \rho_{\text{g}},\ \rho_{\text{ind}}) f(\boldsymbol{\mu}_0|\rho_{\text{g}},\ \rho_{\text{ind}}) d\boldsymbol{\mu}_0 \tag{3.7}$$

Evaluating (3.7) yields,

$$\begin{aligned} p(\mathbf{Y}|\rho_{\text{g}}, \rho_{\text{ind}}) = {} & \frac{|({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T {\boldsymbol{I}^*}^{-1}\mathbf{H})^{-1}|^{1/2}}{(2\pi)^{p/2}|\boldsymbol{I}^*|^{1/2} \times |\boldsymbol{\gamma}_0|^{1/2}} \\ & \times \exp\Big[-\frac{1}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T {\boldsymbol{I}^*}^{-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 + {\boldsymbol{\theta}^*}^T {\boldsymbol{\gamma}_0}^{-1}\boldsymbol{\theta}^* + \mathbf{Y}^T {\boldsymbol{I}^*}^{-1}\mathbf{Y}\Big]\Big] \end{aligned} \tag{3.8}$$

where:

$q = \sum_{i=1}^{g} n_i$, $\quad \boldsymbol{I}^* = \sigma^2 \boldsymbol{I}_{q\times q}$, and $\bar{\boldsymbol{\mu}}_0 = ({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T {\boldsymbol{I}^*}^{-1}\mathbf{H})^{-1}(\mathbf{H}^T {\boldsymbol{I}^*}^{-1}\mathbf{Y} + {\boldsymbol{\gamma}_0}^{-1}\boldsymbol{\theta}^*)$

A proof is provided at the end of this chapter (Proof 3.2). In the course of finding, $p(\mathbf{Y}|\rho_{\text{g}}, \rho_{\text{ind}})$, the posterior distribution of $\boldsymbol{\mu}_0$ is found to be Normal with

[2]Discussed in Chapter 6

$E(\boldsymbol{\mu}_0|\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y}) = \bar{\boldsymbol{\mu}}_0$.

Setting $\mathcal{B}$: Unknown and Constant Observation Variance

When σ^2 is unknown, (3.4) becomes

$$\begin{aligned} P(\mathbf{Y}|\rho_{\text{g}}, \rho_{\text{ind}}) &= \int\int p(\mathbf{Y}|\boldsymbol{\mu}_0, \sigma^2, \rho_g, \rho_{\text{ind}}) f(\boldsymbol{\mu}_0, \sigma^2|\rho_g, \rho_{\text{ind}}) d\boldsymbol{\mu}_0 d\sigma^2 \\ &= \int\int p(\mathbf{Y}|\boldsymbol{\mu}_0, \sigma^2, \rho_g, \rho_{\text{ind}}) f_1(\boldsymbol{\mu}_0|\sigma^2, \rho_g, \rho_{\text{ind}}) f_2(\sigma^2|\rho_g, \rho_{\text{ind}}) d\boldsymbol{\mu}_0 d\sigma^2 \end{aligned} \tag{3.9}$$

Let the component distributions be:

$$p(\mathbf{Y}|\boldsymbol{\mu}_0, \sigma^2, \rho_{\text{g}}, \rho_{\text{ind}}) = \frac{1}{(2\pi)^{q/2}|\Sigma_q|^{1/2}} \exp\Big[-\frac{1}{2}\Big[(\mathbf{Y} - \mathbf{H}\boldsymbol{\mu}_0)^T \Sigma_q^{-1} (\mathbf{Y} - \mathbf{H}\boldsymbol{\mu}_0)\Big]\Big]$$

$$f_1(\boldsymbol{\mu}_0|\sigma^2, \rho_{\text{g}}, \rho_{\text{ind}}) = \frac{1}{(2\pi)^{w/2}|k_0^{-1}\Sigma_w|^{1/2}} \exp\Big[-\frac{1}{2}\Big[(\boldsymbol{\mu}_0 - \boldsymbol{\theta}^*)^T (k_0 \Sigma_w^{-1})(\boldsymbol{\mu}_0 - \boldsymbol{\theta}^*)\Big]\Big]$$

$$f_2(\sigma^2|\rho_{\text{g}}, \rho_{\text{ind}}) = \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)} (\sigma^2)^{-(\nu/2+1)} \exp\Big[-\frac{1}{2}\Big[\nu\sigma_0^2 (\sigma^2)^{-1}\Big]\Big]$$

with $\Sigma_q = \sigma^2 \boldsymbol{I}_q$, $\Sigma_w = \sigma^2 \boldsymbol{I}_w$, $w = ||\boldsymbol{\mu}_0||$, the cardinality of $\boldsymbol{\mu}_0$, and $q = \sum_{i=1}^{g} n_i$ where $\boldsymbol{\theta}^*$, ν, σ_0^2, and k_0 are known, either derived from the data or specified according to the preferences of the modeler. For the above, it can be shown that (3.9) yields

$$P(\mathbf{Y}|\rho_{\text{g}}, \rho_{\text{ind}}) = \frac{|(k_0 \boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}{|k_0^{-1}\boldsymbol{I}_w|^{1/2}} \times \frac{1}{(2\pi)^{p/2}|\boldsymbol{I}_p|^{1/2}} \times \frac{\Gamma(\bar{\nu}/2)}{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}} \times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)}$$

where:

$\bar{\nu} = \nu + p$, $\bar{\nu}\bar{\sigma}^2 = \nu\sigma_0^2 - \bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^T(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta} + \mathbf{Y}^T\mathbf{Y}$, and

$\bar{\boldsymbol{\mu}}_0 = (k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}(\mathbf{H}^T\mathbf{Y} + k_0\boldsymbol{I}_w^{-1}\boldsymbol{\theta})$

A proof is provided at the end of this chapter (Proof 3.3). In the course of finding $p(\mathbf{Y}|\rho_{\text{g}}, \rho_{\text{ind}})$, it is shown that the marginal posterior distribution for $\boldsymbol{\mu}_0$ is a multivariate t-distribution with mean $\bar{\boldsymbol{\mu}}_0$, so therefore $E(\boldsymbol{\mu}_0|\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y}) = \bar{\boldsymbol{\mu}}_0$.

Together, these results allow for a full specification of $P(\rho_{\text{g}},\ \rho_{\text{ind}},\ \mathbf{Y})$ and subsequently $P(\rho_{\text{g}},\ \rho_{\text{ind}}|\mathbf{Y})$ up to a constant of proportionality. Take for example setting $\mathcal{A}$ with known and constant observation variance. Let $P(\rho_{\text{g}})$ and $P(\rho_{\text{ind}})$ have Ewens-Pitman priors where

$$L = \frac{\Gamma(m_{\text{gr}})}{\Gamma(m_{\text{gr}} + g)} \text{ and } K = \frac{\Gamma(m_{\text{ind}})}{\Gamma(m_{\text{ind}} + q)}$$

with m_{gr} and m_{ind} denoting the respective m hyperparameters. Then

$$\begin{aligned}
P(\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y}) &= P(\rho_{\text{g}}, \rho_{\text{ind}}) \times p(\mathbf{Y}|\rho_{\text{g}}, \rho_{\text{ind}}) \\
&= \left[\frac{m_{\text{gr}}^l \Gamma(m_{\text{gr}})}{\Gamma(m_{\text{gr}} + g)} \prod_{t=1}^{l}(n_t - 1)!\right]\left[\frac{m_{\text{ind}}^k \Gamma(m_{\text{ind}})}{\Gamma(m_{\text{ind}} + q)} \prod_{s=1}^{k}(n_s - 1)!\right] \\
&\quad \times \frac{|(\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T\boldsymbol{I}^{*-1}\mathbf{H})^{-1}|^{1/2}}{(2\pi)^{q/2}|\boldsymbol{I}^*|^{1/2} \times |\boldsymbol{\gamma}_0|^{1/2}} \\
&\quad \times \exp\left[-\frac{1}{2}\left[-\bar{\boldsymbol{\mu}}_0^T(\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T\boldsymbol{I}^{*-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^{*T}\boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^* + \mathbf{Y}^T\boldsymbol{I}^{*-1}\mathbf{Y}\right]\right]
\end{aligned}$$

So,

$$P(\rho_{\text{g}}, \rho_{\text{ind}}|\mathbf{Y}) \propto \left[m_{\text{gr}}^l \prod_{t=1}^{l}(n_t - 1)!\right]\left[m_{\text{ind}}^k \prod_{s=1}^{k}(n_s - 1)!\right] \times \frac{|(\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T\boldsymbol{I}^{*-1}\mathbf{H})^{-1}|^{1/2}}{|\boldsymbol{\gamma}_0|^{1/2}}$$

$$\times \exp\Big[-\frac{1}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T(\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T\boldsymbol{I}^{*-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^{*T}\boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^* + \mathbf{Y}^T\boldsymbol{I}^{*-1}\mathbf{Y}\Big]\Big]$$

where the $\mathbf{Y}^T\boldsymbol{I}^{*-1}\mathbf{Y}$ term is retained in the exponent of the above expression for computational scale purposes.

Note how $\mathbf{H}$ tracks through the mathematics and incorporates the individual-level set mean relationships implied by the two-level mean condition into the posterior joint distribution of ρ_{g} and ρ_{ind}. The next chapter details how $\mathbf{H}$ is found for any pair of partitions.

Proofs

Proof 3.1

Letting $c(S) = \frac{(n_S-1)!}{m^{n_S-1}}$ results in a partition distribution claimed in [Cro97] to be equivalent to letting $c(S) = m(n_S - 1)!$. The proof involves a term describing the probability that a specific set S of items is in the partition, called by [Har90, Cro92, Cro97] the *relevance of S*. From [Cro92, Cro97], relevances are uniquely determined while cohesions are not. As such, two different cohesion functions may result in identical relevances for a set S. For this proof, let there be a total of n items. Following [Cro97], the relevances are defined as:

$$r(S) = \frac{c(S)\lambda(S_0-S)}{\lambda(S_0)},$$

where $r(S)$ is the relevance of set S, and S_0 is the set of all n items. $\lambda(S)$ is a function defined by choosing a *specific* element $i \in S$ and computing the recursive

relation:

$$\lambda(S) = \sum_{\{T|i\in T\}} c(T)\lambda(S-T)$$

where $T \subseteq S$ and $\lambda(\emptyset) = 1$. From [Cro92, Cro97], assuming n_S items in set S, if $c(S) = \frac{(n_S-1)!}{m^{n_S-1}}$, then $r(S) = mB(n_s, n+m-n_S)$ where B denotes a Beta function such that $B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$. To complete the proof, since relevances are unique, it suffices to show that changing the cohesion function to $c(S) = m(n_S-1)!$ does not affect the relevances. Thus, letting $c(S) = m(n_S-1)!$, one needs to show the final pair in the following equivalence relation holds:

$$r(S) = \frac{c(S)\lambda(S_0-S)}{\lambda(S_0)} = \frac{m(n_S-1)!\lambda(S_0-S)}{\lambda(S_0)} = \frac{m\Gamma(n_S)\Gamma(n+m-n_S)}{\Gamma(n+m)}$$

Therefore, it is enough to show that

$$\frac{\lambda(S_0-S)}{\lambda(S_0)} = \frac{(n+m-n_S-1)!}{(n+m-1)!}$$

Let S^h be a set with h items. By the definition of λ, since one specific object must be selected from the h objects to "remain in T," there are $h-1$ objects that may be in $S-T$, and thus, as shown in [Cro92],

$$\begin{aligned}
\lambda(S^h) = &\binom{h-1}{0} c(S^h)\lambda(\emptyset) \\
&+ \binom{h-1}{1} c(S^{h-1})\lambda(S^1) \\
&+ \binom{h-1}{2} c(S^{h-2})\lambda(S^2) \\
&+ \dots \\
&+ \binom{h-1}{h-1} c(S^1)\lambda(S^{h-1})
\end{aligned}$$

$$\begin{aligned}
\lambda(S^h) &= \frac{(h-1)!}{0!(h-1)!}m(h-1)! * 1 \\
&+ \frac{(h-1)!}{1!(h-2)!}m(h-2)! * \lambda(S^1) \\
&+ \frac{(h-1)!}{2!(h-3)!}m(h-3)! * \lambda(S^2) \\
&+ \dots \\
&+ \frac{(h-1)!}{(h-1)!(0)!}m(0)! * \lambda(S^{h-1}) \\
&= m(h-1)!\left[1 + \lambda(S^1) + \frac{\lambda(S^2)}{2!} + ... + \frac{\lambda(S^{h-1})}{(h-1)!}\right]
\end{aligned}$$

Then use induction to show that $\lambda(S^h) = \frac{\Gamma(m+h)}{\Gamma(m)}$.

Basis Case:

$$\begin{aligned}
\lambda(S^1) &= c(S^1)\lambda(S^1 - T) \\
&= c(S^1) \qquad \text{Since there is only one element in } S^1 \\
&= m(1-1)! \\
&= m * 1 \\
&= m = \frac{m(m-1)!}{(m-1)!} = \frac{m!}{(m-1)!} = \frac{\Gamma(m+1)}{\Gamma(m)}
\end{aligned}$$

Thus, using $\lambda(S^h) = \frac{\Gamma(m+h)}{\Gamma(m)}$ need to show $\lambda(S^{h+1}) = \frac{\Gamma(m+h+1)}{\Gamma(m)}$.

$$\begin{aligned}
\lambda(S^{h+1}) &= mh!\left[1 + \lambda(S^1) + \frac{\lambda(S^2)}{2!} + ... + \frac{\lambda(S^h)}{(h)!}\right] \\
&= h(m(h-1)!)\left[1 + \lambda(S^1) + \frac{\lambda(S^2)}{2!} + ... + \frac{\lambda(S^h)}{(h)!}\right] \\
&= h\left[\lambda(S^h)\right] + h(m(h-1)!)\left[\frac{\lambda(S^h)}{(h)!}\right] \\
&= h\left[\lambda(S^h)\right] + m\left[\frac{\lambda(S^h)}{(h)!}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\lambda(S^h)}{(h)!}\Big[h+m\Big] \\
&= \frac{\Gamma(m+h)}{\Gamma(m)}(h+m) \\
&= \frac{(m+h-1)!(h+m)}{\Gamma(m)} \\
&= \frac{\Gamma(m+h+1)}{\Gamma(m)}
\end{aligned}
$$

Therefore, since $\lambda(S_0 - S) = \frac{\Gamma(m+n-n_S)}{\Gamma(m)}$ and $\lambda(S_0) = \frac{\Gamma(m+n)}{\Gamma(m)}$,

$$\frac{\lambda(S_0 - S)}{\lambda(S_0)} = \frac{\Gamma(m+n-n_S)}{\Gamma(m+n)}\frac{\Gamma(m)}{\Gamma(m)} = \frac{\Gamma(m+n-n_S)}{\Gamma(m+n)}$$

QED

The proofs below make use of established methods of evaluating conjugate priors for normal data. It is important to note the incorporated **H** matrix, encoding the implications of the two-level mean condition, and how it tracks through the evaluation of the integrals.

Proof 3.2

$p(\mathbf{Y}|\boldsymbol{\mu}_0, \rho_{\mathrm{g}}, \rho_{\mathrm{ind}})f(\boldsymbol{\mu}_0|\rho_{\mathrm{g}}, \rho_{\mathrm{ind}})$

$$
\begin{aligned}
&= \frac{1}{(2\pi)^{q/2}|\boldsymbol{I}^*|^{1/2}}\exp\Big[-\frac{1}{2}\Big[(\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_0)^T\boldsymbol{I}^{*-1}(\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_0)\Big]\Big] \\
&\quad\times\frac{1}{(2\pi)^{w/2}|\boldsymbol{\gamma}_0|^{1/2}}\exp\Big[-\frac{1}{2}\Big[(\boldsymbol{\mu}_0-\boldsymbol{\theta}^*)^T{\boldsymbol{\gamma}_0}^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\theta}^*)\Big]\Big]
\end{aligned}
$$

where: $q = \sum_{i=1}^{g} n_i$, $\quad w = ||\boldsymbol{\mu}_0||$, $\quad \boldsymbol{I}^* = \sigma^2\boldsymbol{I}_{q\times q}$, $\quad \boldsymbol{\theta}^* = \theta\mathbf{1}_w$, $\quad \boldsymbol{\gamma}_0 = \mathrm{c}\boldsymbol{I}_w$. Consider first the exponential terms:

$$(\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_0)^T\boldsymbol{I}^{*-1}(\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_0) + (\boldsymbol{\mu}_0-\boldsymbol{\theta}^*)^T{\boldsymbol{\gamma}_0}^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\theta}^*)$$

Expand the first term:

$$
\begin{aligned}
&= (\mathbf{y}^T \boldsymbol{I}^{*-1} - (H\boldsymbol{\mu}_0)^T \boldsymbol{I}^{*-1})(\mathbf{y} - H\boldsymbol{\mu}_0) \\
&= \mathbf{y}^T \boldsymbol{I}^{*-1}\mathbf{y} - (H\boldsymbol{\mu}_0)^T \boldsymbol{I}^{*-1}\mathbf{y} - \mathbf{y}^T \boldsymbol{I}^{*-1}(H\boldsymbol{\mu}_0) + (H\boldsymbol{\mu}_0)^T \boldsymbol{I}^{*-1}(H\boldsymbol{\mu}_0) \\
&= \mathbf{y}^T \boldsymbol{I}^{*-1}\mathbf{y} - 2(H\boldsymbol{\mu}_0)^T \boldsymbol{I}^{*-1}\mathbf{y} + (H\boldsymbol{\mu}_0)^T \boldsymbol{I}^{*-1}(H\boldsymbol{\mu}_0)
\end{aligned}
$$

Expand the second term:

$$
\begin{aligned}
&= (\boldsymbol{\mu}_0{}^T \boldsymbol{\gamma}_0{}^{-1} - \boldsymbol{\theta}^{*T} \boldsymbol{\gamma}_0{}^{-1})(\boldsymbol{\mu}_0 - \boldsymbol{\theta}^*) \\
&= \boldsymbol{\mu}_0{}^T \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\theta}^{*T} \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_0{}^T \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^* + \boldsymbol{\theta}^{*T} \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^* \\
&= \boldsymbol{\mu}_0{}^T \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\mu}_0 - 2\boldsymbol{\mu}_0{}^T \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^* + \boldsymbol{\theta}^{*T} \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^*
\end{aligned}
$$

Combining term expansions:

$$
\begin{aligned}
&= \boldsymbol{\mu}_0{}^T \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\mu}_0 - 2\boldsymbol{\mu}_0{}^T (\mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{y} + \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^*) \\
&\qquad + \boldsymbol{\theta}^{*T} \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^* + \boldsymbol{\mu}_0^T H^T \boldsymbol{I}^{*-1} H\boldsymbol{\mu}_0 + \mathbf{y}^T \boldsymbol{I}^{*-1}\mathbf{y} \\
&= \boldsymbol{\mu}_0^T (\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{H})\boldsymbol{\mu}_0 - 2\boldsymbol{\mu}_0{}^T (\mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{y} + \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^*) \\
&\qquad + \boldsymbol{\theta}^{*T} \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^* + \mathbf{y}^T \boldsymbol{I}^{*-1}\mathbf{y}
\end{aligned}
$$

Let $\bar{\boldsymbol{\mu}}_0 = (\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{H})^{-1}(\mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{y} + \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^*)$. Then the exponential terms may be rewritten as

$$
\begin{aligned}
&\boldsymbol{\mu}_0^T (\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{H})\boldsymbol{\mu}_0 \\
&\quad - 2\boldsymbol{\mu}_0^T (\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{H})(\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{H})^{-1}(\mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{y} + \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^*) \\
&\quad + \bar{\boldsymbol{\mu}}_0^T (\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 - \bar{\boldsymbol{\mu}}_0^T (\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 \\
&\quad + \boldsymbol{\theta}^{*T} \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^* + \mathbf{y}^T \boldsymbol{I}^{*-1}\mathbf{y}
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{\mu}_0^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})\boldsymbol{\mu}_0 - 2\boldsymbol{\mu}_0^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 \\
&\quad + \bar{\boldsymbol{\mu}}_0^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 - \bar{\boldsymbol{\mu}}_0^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 \\
&\quad + {\boldsymbol{\theta}^*}^T{\boldsymbol{\gamma}_0}^{-1}\boldsymbol{\theta}^* + \mathbf{y}^T{\boldsymbol{I}^*}^{-1}\mathbf{y} \\
&= (\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0) \\
&\quad - \bar{\boldsymbol{\mu}}_0^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 + {\boldsymbol{\theta}^*}^T{\boldsymbol{\gamma}_0}^{-1}\boldsymbol{\theta}^* + \mathbf{y}^T{\boldsymbol{I}^*}^{-1}\mathbf{y}
\end{aligned}
$$

Therefore

$$
\begin{aligned}
p(\mathbf{Y}|\boldsymbol{\mu}_0, \rho_{\mathrm{g}}, \rho_{\mathrm{ind}}) f(\boldsymbol{\mu}_0|\rho_{\mathrm{g}}, \rho_{\mathrm{ind}}) &= \frac{1}{(2\pi)^{q/2}|\boldsymbol{I}^*|^{1/2}} \times \frac{1}{(2\pi)^{w/2}|\boldsymbol{\gamma}_0|^{1/2}} \\
&\quad \times \exp\Big[-\frac{1}{2}\Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)\Big]\Big] \\
&\quad \times \exp\Big[-\frac{1}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 + {\boldsymbol{\theta}^*}^T{\boldsymbol{\gamma}_0}^{-1}\boldsymbol{\theta}^* + \mathbf{y}^T{\boldsymbol{I}^*}^{-1}\mathbf{y}\Big]\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{(2\pi)^{q/2}|\boldsymbol{I}^*|^{1/2}} \times \frac{1}{(2\pi)^{w/2}|\boldsymbol{\gamma}_0|^{1/2}} \\
&\quad \times \frac{1}{(2\pi)^{w/2}|({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})^{-1}|^{1/2}} \times (2\pi)^{w/2}|({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})^{-1}|^{1/2} \\
&\quad \times \exp\Big[-\frac{1}{2}\Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)\Big]\Big] \\
&\quad \times \exp\Big[-\frac{1}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 + {\boldsymbol{\theta}^*}^T{\boldsymbol{\gamma}_0}^{-1}\boldsymbol{\theta}^* + \mathbf{y}^T{\boldsymbol{I}^*}^{-1}\mathbf{y}\Big]\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{(2\pi)^{w/2}|({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})^{-1}|^{1/2}}{(2\pi)^{q/2}|\boldsymbol{I}^*|^{1/2} \times (2\pi)^{w/2}|\boldsymbol{\gamma}_0|^{1/2}} \\
&\quad \times \exp\Big[-\frac{1}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 + {\boldsymbol{\theta}^*}^T{\boldsymbol{\gamma}_0}^{-1}\boldsymbol{\theta}^* + \mathbf{y}^T{\boldsymbol{I}^*}^{-1}\mathbf{y}\Big]\Big] \\
&\quad \times \frac{1}{(2\pi)^{w/2}|({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})^{-1}|^{1/2}} \\
&\quad \times \exp\Big[-\frac{1}{2}\Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T({\boldsymbol{\gamma}_0}^{-1} + \mathbf{H}^T{\boldsymbol{I}^*}^{-1}\mathbf{H})(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)\Big]\Big]
\end{aligned}
$$

So, the posterior distribution of $\boldsymbol{\mu}_0$ is Multivariate Normal:

$$\boldsymbol{\mu}_0|\mathbf{Y}, \rho_{\text{g}}, \rho_{\text{ind}} \sim MVN(\bar{\boldsymbol{\mu}}_0, (\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{H})^{-1})$$

with $E(\boldsymbol{\mu}_0|\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y}) = \bar{\boldsymbol{\mu}}_0$, so $\mathbf{H}E(\boldsymbol{\mu}_0|\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y}) = \mathbf{H}\bar{\boldsymbol{\mu}}_0$.

Therefore, the marginal data distribution is

$$\begin{aligned} p(\mathbf{Y}|\rho_{\text{g}}, \rho_{\text{ind}}) &= \int p(\mathbf{Y}|\boldsymbol{\mu}_0, \rho_{\text{g}}, \rho_{\text{ind}}) f(\boldsymbol{\mu}_0|\rho_{\text{g}}, \rho_{\text{ind}}) d\boldsymbol{\mu}_0 \\ &= \frac{|(\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{H})^{-1}|^{1/2}}{(2\pi)^{q/2}|\boldsymbol{I}^*|^{1/2}|\boldsymbol{\gamma}_0|^{1/2}} \\ &\quad \times \exp\Big[-\frac{1}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T(\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^{*T}\boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^* + \mathbf{y}^T \boldsymbol{I}^{*-1}\mathbf{y}\Big]\Big] \end{aligned}$$

where:

$$q = \textstyle\sum_{i=1}^{g} n_i, \quad w = ||\boldsymbol{\mu}_0||, \quad \boldsymbol{I}^* = \sigma^2 \boldsymbol{I}_{q\times q}, \quad \boldsymbol{\theta}^* = \theta \mathbf{1}_w, \quad \boldsymbol{\gamma}_0 = \text{c}\boldsymbol{I}_w,$$

and

$$\bar{\boldsymbol{\mu}}_0 = (\boldsymbol{\gamma}_0{}^{-1} + \mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{H})^{-1}(\mathbf{H}^T \boldsymbol{I}^{*-1}\mathbf{y} + \boldsymbol{\gamma}_0{}^{-1}\boldsymbol{\theta}^*)$$

QED

Proof 3.3

$$\begin{aligned} P(\mathbf{Y}|\rho_{\text{g}}, \rho_{\text{ind}}) &= \int\int p(\mathbf{Y}|\boldsymbol{\mu}_0, \sigma^2, \rho_g, \rho_{\text{ind}}) f(\boldsymbol{\mu}_0, \sigma^2|\rho_g, \rho_{\text{ind}}) d\boldsymbol{\mu}_0 d\sigma^2 \\ &= \int\int p(\mathbf{Y}|\boldsymbol{\mu}_0, \sigma^2, \rho_g, \rho_{\text{ind}}) f(\boldsymbol{\mu}_0|\sigma^2, \rho_g, \rho_{\text{ind}}) f(\sigma^2|\rho_g, \rho_{\text{ind}}) d\boldsymbol{\mu}_0 d\sigma^2 \end{aligned}$$

with component distributions :

$$p(\mathbf{Y}|\boldsymbol{\mu}_0, \sigma^2, \rho_{\text{g}}, \rho_{\text{ind}}) = \frac{1}{(2\pi)^{q/2}|\Sigma_q|^{1/2}} \exp\Big[-\frac{1}{2}\Big[(\mathbf{Y} - \mathbf{H}\boldsymbol{\mu}_0)^T \Sigma_q^{-1}(\mathbf{Y} - \mathbf{H}\boldsymbol{\mu}_0)\Big]\Big]$$

$$f(\boldsymbol{\mu}_0|\sigma^2,\rho_{\text{g}},\rho_{\text{ind}}) = \frac{1}{(2\pi)^{w/2}|k_0^{-1}\Sigma_w|^{1/2}}\text{exp}\bigg[-\frac{1}{2}\Big[(\boldsymbol{\mu}_0-\boldsymbol{\theta}^*)^T(k_0\Sigma_w^{-1})(\boldsymbol{\mu}_0-\boldsymbol{\theta}^*)\Big]\bigg]$$

$$f(\sigma^2|\rho_{\text{g}},\rho_{\text{ind}}) = \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-(\nu/2+1)}\text{exp}\bigg[-\frac{1}{2}\Big[\nu\sigma_0^2(\sigma^2)^{-1}\Big]\bigg]$$

where $\Sigma_q = \sigma^2\boldsymbol{I}_q$, $\Sigma_w = \sigma^2\boldsymbol{I}_w$, $w = ||\boldsymbol{\mu}_0||$, the cardinality of $\boldsymbol{\mu}_0$, and $q = \sum_{i=1}^{g} n_i$ where $\boldsymbol{\theta}^*$, ν, σ_0^2, and k_0 are known.

Rewrite $p(\mathbf{Y}|\boldsymbol{\mu}_0,\sigma^2,\rho_{\text{g}},\rho_{\text{ind}})$

$$\begin{aligned}
&= \frac{1}{(2\pi)^{q/2}|\sigma^2\boldsymbol{I}_q|^{1/2}}\text{exp}\bigg[-\frac{1}{2}\Big[(\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_0)^T(\sigma^2\boldsymbol{I}_q)^{-1}(\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_0)\Big]\bigg] \\
&= \frac{1}{(2\pi)^{q/2}|\sigma^2\boldsymbol{I}_q|^{1/2}}\text{exp}\bigg[-\frac{(\sigma^2)^{-1}}{2}\Big[(\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_0)^T(\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_0)\Big]\bigg]
\end{aligned}$$

and rewrite $f(\boldsymbol{\mu}_0|\sigma^2,\rho_{\text{g}},\rho_{\text{ind}})$

$$\begin{aligned}
&= \frac{1}{(2\pi)^{w/2}|k_0^{-1}\sigma^2\boldsymbol{I}_w|^{1/2}}\text{exp}\bigg[-\frac{1}{2}\Big[(\boldsymbol{\mu}_0-\boldsymbol{\theta}^*)^T(k_0(\sigma^2\boldsymbol{I}_w)^{-1})(\boldsymbol{\mu}_0-\boldsymbol{\theta}^*)\Big]\bigg] \\
&= \frac{1}{(2\pi)^{w/2}|k_0^{-1}\sigma^2\boldsymbol{I}_w|^{1/2}}\text{exp}\bigg[-\frac{(\sigma^2)^{-1}}{2}\Big[(\boldsymbol{\mu}_0-\boldsymbol{\theta}^*)^T(k_0\boldsymbol{I}_w^{-1})(\boldsymbol{\mu}_0-\boldsymbol{\theta}^*)\Big]\bigg]
\end{aligned}$$

Consider the exponential terms from $p(\mathbf{Y}|\boldsymbol{\mu}_0,\sigma^2,\rho_{\text{g}},\rho_{\text{ind}})$ and $f(\boldsymbol{\mu}_0|\sigma^2,\rho_{\text{g}},\rho_{\text{ind}})$

$$(\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_0)^T(\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_0) + (\boldsymbol{\mu}_0-\boldsymbol{\theta}^*)^T(k_0\boldsymbol{I}_w^{-1})(\boldsymbol{\mu}_0-\boldsymbol{\theta}^*)$$

$$= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\mu}_0^T\mathbf{H}^T\mathbf{y} + \boldsymbol{\mu}_0^T\mathbf{H}^T\mathbf{H}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^T(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\mu}_0 - 2\boldsymbol{\mu}_0^T(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^*$$

$$= \boldsymbol{\mu}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\boldsymbol{\mu}_0 - 2\boldsymbol{\mu}_0^T(\mathbf{H}^T\mathbf{y} + k_0\boldsymbol{I}_w^{-1}\boldsymbol{\theta}^*) + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \mathbf{y}^T\mathbf{y}$$

Let $\bar{\boldsymbol{\mu}}_0 = (k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}(\mathbf{H}^T\mathbf{y} + k_0\boldsymbol{I}_w^{-1}\boldsymbol{\theta}^*)$. Then the exponential terms may be rewritten as:

$$\begin{aligned}
&\boldsymbol{\mu}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\boldsymbol{\mu}_0 - 2\boldsymbol{\mu}^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}(\mathbf{H}^T\mathbf{y} + k_0\boldsymbol{I}_w^{-1}\boldsymbol{\theta}^*) \\
&\qquad + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \mathbf{y}^T\mathbf{y} \\
&= \boldsymbol{\mu}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\boldsymbol{\mu}_0 - 2\boldsymbol{\mu}^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \mathbf{y}^T\mathbf{y} \\
&= \boldsymbol{\mu}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\boldsymbol{\mu}_0 - 2\boldsymbol{\mu}^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 \\
&\qquad - \bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \mathbf{y}^T\mathbf{y} \\
&= (\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_0) - \bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 \\
&\quad + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \mathbf{y}^T\mathbf{y}
\end{aligned}$$

Therefore, the exponentials can be separated and rewritten as:

$$\begin{aligned}
&\exp\Big[-\frac{(\sigma^2)^{-1}}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \mathbf{y}^T\mathbf{y}\Big]\Big] \\
&\qquad \times \exp\Big[-\frac{(\sigma^2)^{-1}}{2}\Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_0)\Big]\Big] \\
&= \exp\Big[-\frac{(\sigma^2)^{-1}}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \mathbf{y}^T\mathbf{y}\Big]\Big] \qquad (3.10) \\
&\qquad \times \exp\Big[-\frac{1}{2}\Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})(\sigma^2)^{-1}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_0)\Big]\Big]
\end{aligned}$$

So multiplying by $\frac{|(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}{|(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}$ and reorganizing (3.10) gives:

$$\begin{aligned}
&\frac{1}{(2\pi)^{q/2}|\sigma^2\boldsymbol{I}_q|^{1/2}} \times \exp\Big[-\frac{(\sigma^2)^{-1}}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \mathbf{y}^T\mathbf{y}\Big]\Big] \\
&\quad \times \frac{1}{(2\pi)^{w/2}|k_0^{-1}\sigma^2\boldsymbol{I}_w|^{1/2}} \times \frac{|(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}{|(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{1/2}} \qquad (3.11)
\end{aligned}$$

$$\times \exp\Big[-\frac{1}{2}\Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})(\sigma^2)^{-1}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_0)\Big]\Big]$$
$$\times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-(\nu/2+1)}\exp\Big[-\frac{1}{2}\Big[\nu\sigma_0^2(\sigma^2)^{-1}\Big]\Big]$$

Since if a is a scalar, $|a\boldsymbol{I}_w| = a^w|\boldsymbol{I}_w|$, (3.11) is equal to

$$\frac{1}{(2\pi)^{w/2}|(\sigma^2)(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}$$
$$\times \exp\Big[-\frac{1}{2}\Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})(\sigma^2)^{-1}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_0)\Big]\Big]$$
$$\times \frac{|(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}{|k_0^{-1}\boldsymbol{I}_w|^{1/2}} \times \frac{1}{(2\pi)^{q/2}|\sigma^2\boldsymbol{I}_q|^{1/2}}$$
$$\times \exp\Big[-\frac{(\sigma^2)^{-1}}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \mathbf{y}^T\mathbf{y}\Big]\Big]$$
$$\times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-(\nu/2+1)}\exp\Big[-\frac{1}{2}\Big[\nu\sigma_0^2(\sigma^2)^{-1}\Big]\Big]$$

Thus, $\boldsymbol{\mu}_0|\sigma^2 \sim MVN(\bar{\boldsymbol{\mu}}_0, (k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}(\sigma^2))$

$$\int \Bigg\{ \frac{1}{(2\pi)^{w/2}|(\sigma^2)(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}$$
$$\times \exp\Big[-\frac{1}{2}\Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})(\sigma^2)^{-1}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_0)\Big]\Big]$$
$$\times \frac{|(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}{|k_0^{-1}\boldsymbol{I}_w|^{1/2}} \times \frac{1}{(2\pi)^{q/2}|\sigma^2\boldsymbol{I}_q|^{1/2}}$$
$$\times \exp\Big[-\frac{(\sigma^2)^{-1}}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \mathbf{y}^T\mathbf{y}\Big]\Big]$$
$$\times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-(\nu/2+1)}\exp\Big[-\frac{1}{2}\Big[\nu\sigma_0^2(\sigma^2)^{-1}\Big]\Big]\Bigg\} d\boldsymbol{\mu}_0$$

$$= \frac{|(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}{|k_0^{-1}\boldsymbol{I}_w|^{1/2}} \times \frac{1}{(2\pi)^{q/2}|\sigma^2\boldsymbol{I}_q|^{1/2}}$$
$$\times \exp\Big[-\frac{(\sigma^2)^{-1}}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \mathbf{y}^T\mathbf{y}\Big]\Big]$$

$$\times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-(\nu/2+1)}\exp\Big[-\frac{1}{2}\Big[\nu\sigma_0^2(\sigma^2)^{-1}\Big]\Big]$$

To integrate out σ^2, consider what remains and rewrite

$$\begin{aligned}
&\frac{1}{(2\pi)^{q/2}|\sigma^2\boldsymbol{I}_q|^{1/2}} \times \exp\Big[-\frac{(\sigma^2)^{-1}}{2}\Big[-\bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1}+\mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0+\boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^*+\mathbf{y}^T\mathbf{y}\Big]\Big] \\
&\quad\times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-(\nu/2+1)}\exp\Big[-\frac{1}{2}\Big[\nu\sigma_0^2(\sigma^2)^{-1}\Big]\Big] \\
&= \frac{1}{(2\pi)^{q/2}|\boldsymbol{I}_q|^{1/2}} \times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-((\nu+p)/2+1)} \\
&\quad\times \exp\Big[\frac{(\sigma^2)^{-1}}{2}\Big[\nu\sigma_0^2-\bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1}+\mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0+\boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^*+\mathbf{y}^T\mathbf{y}\Big]\Big] \\
&= \frac{1}{(2\pi)^{q/2}|\boldsymbol{I}_q|^{1/2}} \times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-((\bar{\nu})/2+1)}\exp\Big[-\frac{(\sigma^2)^{-1}}{2}\bar{\nu}\bar{\sigma}^2\Big]
\end{aligned}$$

where $\bar{\nu}=\nu+q,$ and $\bar{\nu}\bar{\sigma}^2=\nu\sigma_0^2-\bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1}+\mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0+\boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^*+\mathbf{y}^T\mathbf{y}$

Multiply the total expression by $\frac{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}\Gamma(\bar{\nu}/2)}{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}\Gamma(\bar{\nu}/2)}$ and reorganize to find:

$$\begin{aligned}
&\frac{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}\Gamma(\bar{\nu}/2)}{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}\Gamma(\bar{\nu}/2)} \times \frac{|(k_0\boldsymbol{I}_w^{-1}+\mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}{|k_0^{-1}\boldsymbol{I}_w|^{1/2}} \times \frac{1}{(2\pi)^{q/2}|\boldsymbol{I}_q|^{1/2}} \\
&\quad\times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-((\bar{\nu})/2+1)}\exp\Big[-\frac{(\sigma^2)^{-1}}{2}\bar{\nu}\bar{\sigma}^2\Big]
\end{aligned}$$

$$\begin{aligned}
&= \frac{|(k_0\boldsymbol{I}_w^{-1}+\mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}{|k_0^{-1}\boldsymbol{I}_w|^{1/2}} \times \frac{1}{(2\pi)^{q/2}|\boldsymbol{I}_q|^{1/2}} \times \frac{\Gamma(\bar{\nu}/2)}{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}} \times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \\
&\quad\times \frac{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}}{\Gamma(\bar{\nu}/2)}(\sigma^2)^{-((\bar{\nu})/2+1)}\exp\Big[-\frac{(\sigma^2)^{-1}}{2}\bar{\nu}\bar{\sigma}^2\Big]
\end{aligned}$$

Therefore, $\sigma^2 \sim$ scale-inverse $\chi^2(\bar{\nu},\bar{\sigma}^2)$ such that

$$\begin{aligned}
&\int \frac{|(k_0\boldsymbol{I}_w^{-1}+\mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}{|k_0^{-1}\boldsymbol{I}_w|^{1/2}} \times \frac{1}{(2\pi)^{q/2}|\boldsymbol{I}_q|^{1/2}} \times \frac{\Gamma(\bar{\nu}/2)}{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}} \times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \\
&\quad\times \frac{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}}{\Gamma(\bar{\nu}/2)}(\sigma^2)^{-((\bar{\nu})/2+1)}\exp\Big[-\frac{(\sigma^2)^{-1}}{2}\bar{\nu}\bar{\sigma}^2\Big]d\sigma^2
\end{aligned}$$

$$= \frac{|(k_0 \boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}{|k_0^{-1}\boldsymbol{I}_w|^{1/2}} \times \frac{1}{(2\pi)^{q/2}|\boldsymbol{I}_q|^{1/2}} \times \frac{\Gamma(\bar{\nu}/2)}{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}} \times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)}$$

where $\bar{\nu} = \nu + q$, $\bar{\nu}\bar{\sigma}^2 = \nu\sigma_0^2 - \bar{\boldsymbol{\mu}}_0^T(k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})\bar{\boldsymbol{\mu}}_0 + \boldsymbol{\theta}^{*T}(k_0\boldsymbol{I}_w^{-1})\boldsymbol{\theta}^* + \mathbf{y}^T\mathbf{y}$, and $\bar{\boldsymbol{\mu}}_0 = (k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}(\mathbf{H}^T\mathbf{y} + k_0\boldsymbol{I}_w^{-1}\boldsymbol{\theta}^*)$

Therefore, $p(Y|\rho_{\text{g}}, \rho_{\text{ind}})$

$$= \frac{|(k_0 \boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}{|k_0^{-1}\boldsymbol{I}_w|^{1/2}} \times \frac{1}{(2\pi)^{q/2}|\boldsymbol{I}_q|^{1/2}} \times \frac{\Gamma(\bar{\nu}/2)}{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}} \times \frac{(\sigma_0^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)}$$

As in the known variance setting, it is necessary to compute the posterior marginal distribution of for $\boldsymbol{\mu}_0$. This can be derived by noting the posterior joint distribution of $\boldsymbol{\mu}_0$ and σ^2:

$$\begin{aligned} f(\boldsymbol{\mu}_0, \sigma^2|\mathbf{Y}, \rho_{\text{g}}, \rho_{\text{ind}}) &= f(\boldsymbol{\mu}_0|\sigma^2, \mathbf{Y}, \rho_{\text{g}}, \rho_{\text{ind}}) f(\sigma^2|\mathbf{Y}, \rho_{\text{g}}, \rho_{\text{ind}}) \\ &= MVN(\bar{\boldsymbol{\mu}}_0, (k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}(\sigma^2)) \times \text{ scale-inverse } \chi^2(\bar{\nu}, \bar{\sigma}^2) \\ &= MVN(\bar{\boldsymbol{\mu}}_0, (k_0\boldsymbol{I}_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}(\sigma^2)) \times IG(\frac{\bar{\nu}}{2}, \frac{\bar{\nu}\bar{\sigma}^2}{2}) \end{aligned}$$

So, $f(\boldsymbol{\mu}_0, \sigma^2|\mathbf{Y}, \rho_{\text{g}}, \rho_{\text{ind}})$

$$\begin{aligned} &= \frac{1}{(2\pi)^{w/2}|(\sigma^2)(k_0 I_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}|^{-1/2}} \\ &\quad \times \exp\Big[-\frac{1}{2}\Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T(k_0 I_w^{-1} + \mathbf{H}^T\mathbf{H})(\sigma^2)^{(-1)}(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)\Big]\Big] \\ &\quad \times \frac{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}}{\Gamma(\bar{\nu}/2)}(\sigma^2)^{-(\bar{\nu}/2+1)}\exp\Big[-\frac{(\sigma^2)^{-1}}{2}\bar{\sigma}^2\bar{\nu}\Big] \end{aligned}$$

Let $V_w = (k_0 I_w^{-1} + \mathbf{H}^T\mathbf{H})^{-1}$ and rewrite:

$$= \frac{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}}{(2\pi)^{w/2}|\sigma^2 V_w|^{-1/2}\Gamma(\bar{\nu}/2)}(\sigma^2)^{-(\bar{\nu}/2+1)} \times \exp\Bigg[-\frac{(\sigma^2)^{-1}}{2}\Big[\bar{\nu}\bar{\sigma}^2 + \Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T V^{-1}(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)\Big]\Big]\Bigg]$$

$$= \frac{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}}{(2\pi)^{w/2}|V_w|^{-1/2}\Gamma(\bar{\nu}/2)}(\sigma^2)^{-(\bar{\nu}/2+w/2+1)} \times \exp\Bigg[-\frac{(\sigma^2)^{-1}}{2}\Big[\bar{\nu}\bar{\sigma}^2 + \Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T V^{-1}(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)\Big]\Big]\Bigg]$$

To obtain the posterior marginal distribution for $\boldsymbol{\mu}_0$, integrate out σ^2. Omitting leading constants for now,

$$f(\boldsymbol{\mu}_0|\mathbf{Y}, \rho_{\text{g}}, \rho_{\text{ind}}) \propto \int_0^\infty (\sigma^2)^{-(\bar{\nu}/2+w/2+1)}\exp\Bigg[-\frac{(\sigma^2)^{-1}}{2}\Big[\bar{\nu}\bar{\sigma}^2 + \Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T V^{-1}(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)\Big]\Big]\Bigg]d\sigma^2$$

Making the substitution $\sigma^2 = \frac{A}{2z}$ where $A = \bar{\nu}\bar{\sigma}^2 + \Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T V^{-1}(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)\Big]$ which implies $z = \frac{A}{2z}$ and $d\sigma^2 = \frac{-A}{2z^2}dz$. Therefore,

$$\begin{aligned} f(\boldsymbol{\mu}_0|\mathbf{Y}, \rho_{\text{g}}, \rho_{\text{ind}}) &\propto \int_0^\infty \left(\frac{A}{2z}\right)^{-(\bar{\nu}/2+w/2+1)} \exp\Big[-z\Big]\left(\frac{-A}{2z^2}\right)dz \\ &\propto \int_0^\infty \left(\frac{A}{2z}\right)^{-(\bar{\nu}/2+w/2+1)} \left(\frac{A}{2}\right) z^{-2}\exp\Big[-z\Big]dz \\ &\propto \left(\frac{A}{2}\right)^{-(\bar{\nu}/2+w/2)} \int_0^\infty z^{-(\bar{\nu}/2+w/2)-1}\exp\Big[-z\Big]dz \end{aligned}$$

The integrand is the kernel of a Gamma$(\bar{\nu}/2 + w/2, 1)$, thus $f(\boldsymbol{\mu}_0|\mathbf{Y}, \rho_{\text{g}}, \rho_{\text{ind}})$

$$\propto \Gamma\left(\frac{\bar{\nu}}{2} + \frac{w}{2}\right)\left[\frac{A}{2}\right]^{-(\bar{\nu}/2+w/2)}$$

$$\propto \Gamma\left(\frac{\bar{\nu}}{2} + \frac{w}{2}\right)\left[\frac{\bar{\nu}\bar{\sigma}^2}{2} + \frac{\Big[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T V^{-1}(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)\Big]}{2}\right]^{-(\bar{\nu}/2+w/2)}$$

$$
\begin{aligned}
&\propto \Gamma\left(\frac{\bar{\nu}}{2}+\frac{w}{2}\right)\left[\bar{\nu}\bar{\sigma}^2+\left[(\boldsymbol{\mu}_0-\bar{\boldsymbol{\mu}}_0)^T V^{-1}\boldsymbol{\mu}_0-\bar{\boldsymbol{\mu}}_0)\right]\right]^{-(\bar{\nu}/2+w/2)} \\
&\propto \Gamma\left(\frac{\bar{\nu}}{2}+\frac{w}{2}\right)\left(\frac{1}{2}\right)^{-(\bar{\nu}/2+w/2)}\left[\bar{\nu}\bar{\sigma}^2+\left[(\boldsymbol{\mu}_0-\bar{\boldsymbol{\mu}}_0)^T V^{-1}\boldsymbol{\mu}_0-\bar{\boldsymbol{\mu}}_0)\right]\right]^{-(\bar{\nu}/2+w/2)} \\
&\propto \Gamma\left(\frac{\bar{\nu}}{2}+\frac{w}{2}\right)\left(\frac{1}{2}\right)^{-(\bar{\nu}/2+w/2)}\left[\bar{\nu}\bar{\sigma}^2\left\{1+\left[(\boldsymbol{\mu}_0-\bar{\boldsymbol{\mu}}_0)^T(\bar{\nu}\bar{\sigma}^2 V)^{-1}\boldsymbol{\mu}_0-\bar{\boldsymbol{\mu}}_0)\right]\right\}\right]^{-(\bar{\nu}/2+w/2)} \\
&\propto \Gamma\left(\frac{\bar{\nu}}{2}+\frac{w}{2}\right)\left(\frac{1}{2}\right)^{-(\bar{\nu}/2+w/2)}\left[\bar{\nu}\bar{\sigma}^2\left\{1+\frac{\left[(\boldsymbol{\mu}_0-\bar{\boldsymbol{\mu}}_0)^T\left(\frac{\bar{\nu}\bar{\sigma}^2}{\bar{\nu}}V\right)^{-1}\boldsymbol{\mu}_0-\bar{\boldsymbol{\mu}}_0)\right]}{\bar{\nu}}\right\}\right]^{-(\bar{\nu}/2+w/2)} \\
&\propto \Gamma\left(\frac{\bar{\nu}}{2}+\frac{w}{2}\right)\left(\frac{1}{2}\right)^{-(\bar{\nu}/2+w/2)}\left[\bar{\nu}\bar{\sigma}^2\right]^{-(\bar{\nu}/2+w/2)} \\
&\qquad\times\left\{1+\frac{\left[(\boldsymbol{\mu}_0-\bar{\boldsymbol{\mu}}_0)^T\left(\frac{\bar{\nu}\bar{\sigma}^2}{\bar{\nu}}V\right)^{-1}\boldsymbol{\mu}_0-\bar{\boldsymbol{\mu}}_0)\right]}{\bar{\nu}}\right\}^{-(\bar{\nu}/2+w/2)}
\end{aligned}
$$

Reintroducing and manipulating the leading terms:

$$
\frac{(\bar{\sigma}^2\bar{\nu}/2)^{\bar{\nu}/2}}{(2\pi)^{w/2}|V_w|^{-1/2}\Gamma(\bar{\nu}/2)}\Gamma\left(\frac{\bar{\nu}}{2}+\frac{w}{2}\right)\left(\frac{1}{2}\right)^{-(\bar{\nu}/2+w/2)}\left[\bar{\nu}\bar{\sigma}^2\right]^{-(\bar{\nu}/2+w/2)}
$$

$$
\begin{aligned}
&= \frac{(\bar{\sigma}^2\bar{\nu}/2)^{-w/2}\Gamma\left(\frac{\bar{\nu}}{2}+\frac{w}{2}\right)}{(2\pi)^{w/2}|V_w|^{-1/2}\Gamma(\bar{\nu}/2)} \\
&= \frac{\Gamma\left(\frac{\bar{\nu}+w}{2}\right)}{(2\pi)^{w/2}|\frac{\bar{\sigma}^2\bar{\nu}}{2}V_w|^{-1/2}\Gamma(\bar{\nu}/2)} \\
&= \frac{\Gamma\left(\frac{\bar{\nu}+w}{2}\right)}{\Gamma(\bar{\nu}/2)(2)^{w/2}(\frac{1}{2})^{w/2}(\pi)^{w/2}|\bar{\sigma}^2\bar{\nu}V_w|^{-1/2}}
\end{aligned}
$$

$$= \frac{\Gamma\left(\frac{\bar{\nu}+w}{2}\right)}{\Gamma(\bar{\nu}/2)(\bar{\nu})^{w/2}(\pi)^{w/2}|\frac{\bar{\sigma}^2\bar{\nu}}{\bar{\nu}}V_w|^{-1/2}}$$

Which together yields the density function for a multivariate t-distribution with parameters $\bar{\boldsymbol{\mu}}_0$ and $\frac{\bar{\nu}\bar{\sigma}^2}{\bar{\nu}}V$:

$$f(\boldsymbol{\mu}_0|\mathbf{Y}, \rho_{\mathrm{g}}, \rho_{\mathrm{ind}})$$
$$= \frac{\Gamma\left(\frac{\bar{\nu}+w}{2}\right)}{\Gamma(\bar{\nu}/2)(\bar{\nu})^{w/2}(\pi)^{w/2}|\frac{\bar{\sigma}^2\bar{\nu}}{\bar{\nu}}V_w|^{-1/2}}\left\{1 + \frac{\left[(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)^T\left(\frac{\bar{\nu}\bar{\sigma}^2}{\bar{\nu}}V\right)^{-1}\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_0)\right]}{\bar{\nu}}\right\}^{-(\bar{\nu}/2+w/2)}$$

for which the expected value, $E(\boldsymbol{\mu}_0|\mathbf{Y}, \rho_{\mathrm{g}}, \rho_{\mathrm{ind}})$, is $\bar{\boldsymbol{\mu}}_0$.

QED

4 Encoding the Two-Level Mean Condition

The Encoding

As noted above, if there are k sets (or equivalently clusters) in the individual-level partition, then the mean of each item's observation distribution is one of k individual-level set means, $\boldsymbol{\mu}_{\text{var}} = \mu^{S_1}, \mu^{S_2}, ..., \mu^{S_k}$. The two-level structure implies a potential relationship between certain individual-level set means. Restating definitions from Chapter 2, for any pair of partitions :

The vector of observation means: $\boldsymbol{\mu} = \left[\mu_{11}, \mu_{12}, ..., \mu_{1n_1}, \mu_{21}, ..., \mu_{gn_g}\right]$,

The vector of individual-level set means: $\boldsymbol{\mu}_{\text{var}} = \left[\mu^{S_1}, \mu^{S_2}, ..., \mu^{S_k}\right]$,

The vector of free set means: $\boldsymbol{\mu}_0 \subseteq \boldsymbol{\mu}_{\text{var}}$,

The vector of linearly related set means: $\boldsymbol{\mu}_{\text{rel}} = \boldsymbol{\mu}_{\text{var}} \setminus \boldsymbol{\mu}_0$

Again, a mapping matrix $\mathbf{H}$ encodes the linear relationships between the set means implied by the two-level mean condition and partitions ρ_{g} and ρ_{ind} such that $\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\mu}_0$. The construction of $\mathbf{H}$ follows below.

For any pair of partitions, ρ_{g} and ρ_{ind}, define the matrices $\mathbf{A}$, $\mathbf{D}$, and $\mathbf{E}$ as follows.

$\mathbf{A}$ is the matrix formed by finding the difference in the average number of each individual-level set mean represented in groups mapped to the same group-level set. Each row corresponds to an implied group-level set mean equality. Each column corresponds to an individual-level set mean. The unique reduced row echelon form (rref) of $\mathbf{A}$ allows for the identification of the free and linearly related set means, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_{\text{rel}}$. $\mathbf{A}$ may be derived through the construction of two other matrices, $\mathbf{G}$ and $\boldsymbol{\Delta}$.

The matrix $\mathbf{G}$ may be identified and determined by ρ_{g} and ρ_{ind} such that $\boldsymbol{\mu} = \mathbf{G}^T\boldsymbol{\mu}_{\text{var}}$. $\mathbf{G}$, operating similarly to k_1, maps items to individual-level sets.

$$\mathbf{G}\Big[, n^*\Big] = \Big[0_1, 0_2, ..., 0_{k'-1}, 1_{k'}, 0_{k'+1}, ..., 0_k\Big]^T,$$

where k' is such that $k_1(\psi(n^*)) = S_{k'}$, where $\psi(n^*) \in \Omega_1$, $\psi(n^*) \neq \psi(n^{*'})$ if $n^* \neq n^{*'}$, and $n^* = 1, ..., \sum_{i=1}^{g} n_i$ such that $\mathbf{G}$ is a $k \times \sum_{i=1}^{g} n_i$ matrix.[1]

Define $\boldsymbol{\Delta}$ to be a $\sum_{i=1}^{g} n_i = q$ column matrix defined by the group equivalences implied by ρ_{g}. For each pair of group indices[2] $(i, i'), i \neq i'$, such that $k_2(i) = k_2(i')$, a row ω in $\boldsymbol{\Delta}$ is defined such that:

$$\begin{cases} \boldsymbol{\Delta}[\omega, \psi(n^*)] = 1 & \text{if item } \psi(n^*) \text{ has group index } i \\ \boldsymbol{\Delta}[\omega, \psi(n^{*'})] = -1 & \text{if item } \psi(n^{*'}) \text{ has group index } i' \\ 0 & \text{otherwise} \end{cases}$$

$\mathbf{A}$ may then be defined as $\mathbf{A} = \boldsymbol{\Delta}\mathbf{G}^T$.

[1] It is helpful to realize that ψ is a mapping function of $1, ..., \sum_{i=1}^{g} n_i$ to $\Omega_1 = \{11, 12, ..., 1n_1, 21, ..., gn_g\}$ for record keeping purposes. A simple version of this function is $\psi(1) = 11, \psi(2) = 12, ..., \psi(n_1) = 1n_1, \psi(n_1 + 1) = 21, ..., \psi(\sum_{i=1}^{g} n_i) = gn_g$

[2] The total number of rows need not be greater than the $g - l$ implied group-level set mean equalities. For example, if $\rho_{\text{g}} = \big\{T_1 = \{1, 2, 3\}\big\}$, only two rows are necessary: one for the relation of group indices 1 and 2 and another for 1 and 3. The mean equivalence for indices 2 and 3 is implied by the others and may be omitted.

Since in rref, any element that is the *leading 1* of its row also requires the elements of its column to all equal 0, the individual-level set mean corresponding to that element's column may be defined in terms of the set means corresponding to the non-zero elements in its row. Therefore, any individual-level set mean corresponding to a column in the rref of $\mathbf{A}$ which possesses a leading 1, must be a linear combination of other set means, and is thus a linearly related set mean. $\mathbf{D}$ and $\mathbf{E}$ are matrices determined by $\mathbf{A}$, $\boldsymbol{\mu}_0$, and $\boldsymbol{\mu}_{\text{rel}}$ (and thus in a broader sense by ρ_{g} and ρ_{ind}) such that $\boldsymbol{\mu}_{\text{var}} = \mathbf{D}\boldsymbol{\mu}_{\text{rel}} + \mathbf{E}\boldsymbol{\mu}_0$. This is the crucial relationship to ensure the two-level mean condition holds.

Lemma 4.1: The vector of set means $\boldsymbol{\mu}_{\text{var}}$ may be expressed as the linear combination of $\mathbf{A}$, $\mathbf{D}$, $\mathbf{E}$, and $\boldsymbol{\mu}_0$: $\boldsymbol{\mu}_{\text{var}} = \mathbf{D}\Big(\mathbf{AD}\Big)^{-1}\Big(- \mathbf{AE}\boldsymbol{\mu}_0\Big) + \mathbf{E}\boldsymbol{\mu}_0$.

Proof 4.1:

Since $\mathbf{A}\boldsymbol{\mu}_{\text{var}} = \mathbf{b}$ (by construction, in this formulation $\mathbf{b}$ will always equal a vector of zeroes):

$$
\begin{aligned}
\implies \mathbf{A}\Big(\mathbf{D}\boldsymbol{\mu}_{\text{rel}} + \mathbf{E}\boldsymbol{\mu}_0\Big) &= \mathbf{b} \\
\mathbf{AD}\boldsymbol{\mu}_{\text{rel}} + \mathbf{AE}\boldsymbol{\mu}_0 &= \mathbf{b} \\
\mathbf{AD}\boldsymbol{\mu}_{\text{rel}} &= \mathbf{b} - \mathbf{AE}\boldsymbol{\mu}_0 \\
\boldsymbol{\mu}_{\text{rel}} &= \Big(\mathbf{AD}\Big)^{-1}\Big(\mathbf{b} - \mathbf{AE}\boldsymbol{\mu}_0\Big)
\end{aligned}
$$

Since $\boldsymbol{\mu}_{\text{var}} = \mathbf{D}\boldsymbol{\mu}_{\text{rel}} + \mathbf{E}\boldsymbol{\mu}_0$,

$$
\implies \boldsymbol{\mu}_{\text{var}} = \mathbf{D}\Big(\mathbf{AD}\Big)^{-1}\Big(\mathbf{b} - \mathbf{AE}\boldsymbol{\mu}_0\Big) + \mathbf{E}\boldsymbol{\mu}_0
$$

Since each row of $\mathbf{A}$ is determined by taking the difference in the average number of

set means (coefficients) in groups mapped to the same cluster and hence have *equal* group means. Thus also, $\mathbf{D}\Big(\mathbf{AD}\Big)^{-1}\mathbf{b} = \mathbf{0}_{||\boldsymbol{\mu}_{\text{var}}||}$, where $||\boldsymbol{\mu}_{\text{var}}||$ is the cardinality of $\boldsymbol{\mu}_{\text{var}}$ and equivalently the number of individual level sets . Therefore:

$$\implies \boldsymbol{\mu}_{\text{var}} = \mathbf{D}\Big(\mathbf{AD}\Big)^{-1}\Big(- \mathbf{AE}\boldsymbol{\mu}_0\Big) + \mathbf{E}\boldsymbol{\mu}_0$$

QED

Lemma 4.2: The definition of $\boldsymbol{\mu}_{\text{var}}$, in terms of $\boldsymbol{\mu}_0$, and $\mathbf{G}$ allow $\boldsymbol{\mu}$ to be defined directly through the free individual-level set means as $\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\mu}_0$ where

$$\boldsymbol{H} = \mathbf{G}^T\Bigg[- \mathbf{D}\Big(\mathbf{AD}\Big)^{-1}\mathbf{AE} + \mathbf{E}\Bigg]$$

Proof 4.2

$$\begin{aligned}
\boldsymbol{\mu} &= \mathbf{G}^T\boldsymbol{\mu}_{\text{var}} \\
&= \mathbf{G}^T\Bigg[\mathbf{D}\Big(\mathbf{AD}\Big)^{-1}\Big(- \mathbf{AE}\boldsymbol{\mu}_0\Big) + \mathbf{E}\boldsymbol{\mu}_0\Bigg] \\
&= \mathbf{G}^T\Bigg[- \mathbf{D}\Big(\mathbf{AD}\Big)^{-1}\mathbf{AE} + \mathbf{E}\Bigg]\boldsymbol{\mu}_0 \\
&= \mathbf{H}\boldsymbol{\mu}_0
\end{aligned}$$

QED

An Example

The context of an illustrative example is helpful. Let there be three groups each with six individual members:

$$\text{Group 1: } 11, 12, 13, 14, 15, 16$$

$$\text{Group 2: } 21, 22, 23, 24, 25, 26$$

$$\text{Group 3: } 31, 32, 33, 34, 35, 36$$

Let the group and individual-level partitions be:

$$\rho_{\text{g}} = \Big\{\{1, 2, 3\}\Big\}$$

$$\rho_{\text{ind}} = \Big\{\{11, 31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36\}\Big\}$$

alternatively written:

$$\rho_{\text{g}} = \{T_1\},\ T_1 = \{1, 2, 3\}$$

$$\rho_{\text{ind}} = \Big\{S_1, S_2, ..., S_6\Big\},$$

$$S_1 = \{11, 31, 32, 33\},\ S_2 = \{12\},\ S_3 = \{13, 21, 22, 23\},$$

$$S_4 = \{14, 24, 25, 26\},\ S_5 = \{15\},\ S_6 = \{16, 34, 35, 36\}$$

which implies:

$$\mu_{11} = \mu^{S_1},\ \mu_{12} = \mu^{S_2},\ \mu_{13} = \mu^{S_3},\ \mu_{14} = \mu^{S_4},\ \mu_{15} = \mu^{S_5},\ \mu_{16} = \mu^{S_6},$$

$$\mu_{21} = \mu^{S_3},\ \mu_{22} = \mu^{S_3},\ \mu_{23} = \mu^{S_3},\ \mu_{24} = \mu^{S_4},\ \mu_{25} = \mu^{S_4},\ \mu_{26} = \mu^{S_4},$$

$$\mu_{31} = \mu^{S_1},\ \mu_{32} = \mu^{S_1},\ \mu_{33} = \mu^{S_1},\ \mu_{34} = \mu^{S_6},\ \mu_{35} = \mu^{S_6},\ \mu_{36} = \mu^{S_6},$$

and since all three group indices are mapped to the same group-level partition set T_1

$$\begin{aligned} n_1^{-1}\sum_{j=1}^{n_1}\mu_{1j} = \mu_1 = \mu_2 = n_2^{-1}\sum_{j=1}^{n_2}\mu_{2j} \\ = \mu_3 = n_3^{-1}\sum_{j=1}^{n_3}\mu_{3j} \end{aligned}$$

The equality between μ_1 and μ_2 implies that

$$\frac{\mu^{S_1}+\mu^{S_2}+\mu^{S_3}+\mu^{S_4}+\mu^{S_5}+\mu^{S_6}}{6} = \frac{3\mu^{S_3}+3\mu^{S_4}}{6}$$

which implies that

$$\begin{aligned} &\mu^{S_1}+\mu^{S_2}+\mu^{S_3}+\mu^{S_4}+\mu^{S_5}+\mu^{S_6}-3\mu^{S_3}-3\mu^{S_4}=0 \\ &\implies \mu^{S_1}+\mu^{S_2}+-2\mu^{S_3}+-2\mu^{S_4}+\mu^{S_5}+\mu^{S_6}=0 \end{aligned} \tag{4.1}$$

and the equality between μ_1 and μ_3 implies that

$$\frac{\mu^{S_1}+\mu^{S_2}+\mu^{S_3}+\mu^{S_4}+\mu^{S_5}+\mu^{S_6}}{6} = \frac{3\mu^{S_1}+3\mu^{S_6}}{6}$$

which implies that

$$\begin{aligned} &\mu^{S_1}+\mu^{S_2}+\mu^{S_3}+\mu^{S_4}+\mu^{S_5}+\mu^{S_6}-3\mu^{S_1}-3\mu^{S_6}=0 \\ &\implies -2\mu^{S_1}+\mu^{S_2}+\mu^{S_3}+\mu^{S_4}+\mu^{S_5}+-2\mu^{S_6}=0 \end{aligned} \tag{4.2}$$

The system of equations (4.1) and (4.2) can be expressed in matrix form as:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & -2 & -2 & 1 & 1 \\ -2 & 1 & 1 & 1 & 1 & -2 \end{bmatrix}$$

which in rref is:

$$\begin{bmatrix} 1 & 0 & -1 & -1 & 0 & 1 \\ 0 & 1 & -1 & -1 & 1 & 0 \end{bmatrix}$$

Using the construction from above, first note that $\mathbf{A}\boldsymbol{\mu}_{\text{var}} = \mathbf{b}$, where $\boldsymbol{\mu}_{\text{var}}^T = \begin{bmatrix} \mu^{S_1} & \mu^{S_2} & \mu^{S_3} & \mu^{S_4} \mu^{S_5} & \mu^{S_6} \end{bmatrix}$ and $\mathbf{b}^T = \begin{bmatrix} 0 & 0 \end{bmatrix}$.

$\mathbf{A}$ can be confirmed directly using $\mathbf{G}$ and $\boldsymbol{\Delta}$. In the present example,

$$\begin{aligned} \boldsymbol{\mu} &= \begin{bmatrix} \mu^{S_1} & \mu^{S_2} & \mu^{S_3} & \mu^{S_4} & \mu^{S_5} & \mu^{S_6} & \mu^{S_3} & \mu^{S_3} & \mu^{S_3} & \mu^{S_4} & \mu^{S_4} & \mu^{S_4} & \mu^{S_1} & \mu^{S_1} & \mu^{S_1} & \mu^{S_6} & \mu^{S_6} & \mu^{S_6} \end{bmatrix}^T \\ &= \mathbf{G}^T \boldsymbol{\mu}_{\text{var}} \\ &= \mathbf{G}^T \begin{bmatrix} \mu^{S_1} & \mu^{S_2} & \mu^{S_3} & \mu^{S_4} & \mu^{S_5} & \mu^{S_6} \end{bmatrix}^T \end{aligned}$$

where

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

For $\boldsymbol{\Delta}$, as noted above there are three equalities implied by ρ_{g}: $\mu_1 = \mu_2$, $\mu_1 = \mu_3$, and $\mu_2 = \mu_3$. The first two relations are sufficient to imply the third. Accordingly, $\boldsymbol{\Delta}$ may be expressed as:

$$\boldsymbol{\Delta} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}$$

Thus,

$$\mathbf{\Delta G}^T = \begin{bmatrix} 1 & 1 & -2 & -2 & 1 & 1 \\ -2 & 1 & 1 & 1 & 1 & -2 \end{bmatrix}$$

which again in rref is

$$\begin{bmatrix} 1 & 0 & -1 & -1 & 0 & 1 \\ 0 & 1 & -1 & -1 & 1 & 0 \end{bmatrix}$$

As each column of the rref of $\mathbf{A}$ corresponds to a set mean, the columns with leading 1's in each row correspond to the set means which are linearly related to the set means corresponding to the other columns. Therefore, μ^{S_1} and μ^{S_2} are linearly related to $\mu^{S_3},\ \mu^{S_4},\ \mu^{S_5}$, and μ^{S_6}:

$$\begin{bmatrix} \textcircled{1} & 0 & -1 & -1 & 0 & 1 \\ 0 & \textcircled{1} & -1 & -1 & 1 & 0 \end{bmatrix}$$

This also implies that $\mu^{S_3},\ \mu^{S_4},\ \mu^{S_5}$, and μ^{S_6} are free means. $\boldsymbol{\mu}_{\text{var}}$ may be expressed as a sum of free and linearly related means:

$$\boldsymbol{\mu}_{\text{var}} = \mathbf{D}\boldsymbol{\mu}_{\text{rel}} + \mathbf{E}\boldsymbol{\mu}_0$$

Here $\boldsymbol{\mu}_{\text{rel}}^T = \begin{bmatrix} \mu^{S_1} & \mu^{S_2} \end{bmatrix}$ and $\boldsymbol{\mu}_0^T = \begin{bmatrix} \mu^{S_3} & \mu^{S_4} & \mu^{S_5} & \mu^{S_6} \end{bmatrix}$, and $\mathbf{D}$ and $\mathbf{E}$ are:

$$D = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad E = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

So:

$$\begin{bmatrix} \mu^{S_1} \\ \mu^{S_2} \\ \mu^{S_3} \\ \mu^{S_4} \\ \mu^{S_5} \\ \mu^{S_6} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mu^{S_1} \\ \mu^{S_2} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu^{S_3} \\ \mu^{S_4} \\ \mu^{S_5} \\ \mu^{S_6} \end{bmatrix}$$

So calculate:

$$(AD)^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad D(AD)^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad D(AD)^{-1}AE = \begin{bmatrix} -1 & -1 & 0 & 1 \\ -1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Thus,

$$\begin{bmatrix} \mu^{S_1} \\ \mu^{S_2} \\ \mu^{S_3} \\ \mu^{S_4} \\ \mu^{S_5} \\ \mu^{S_6} \end{bmatrix} = - \begin{bmatrix} -1 & -1 & 0 & 1 \\ -1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu^{S_3} \\ \mu^{S_4} \\ \mu^{S_6} \\ \mu^{S_6} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu^{S_3} \\ \mu^{S_4} \\ \mu^{S_5} \\ \mu^{S_6} \end{bmatrix}$$

$$\begin{bmatrix} \mu^{S_1} \\ \mu^{S_2} \\ \mu^{S_3} \\ \mu^{S_4} \\ \mu^{S_5} \\ \mu^{S_6} \end{bmatrix} = \begin{bmatrix} \mu^{S_3} + \mu^{S_4} - \mu^{S_6} \\ \mu^{S_3} + \mu^{S_4} - \mu^{S_5} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \mu^{S_3} \\ \mu^{S_4} \\ \mu^{S_5} \\ \mu^{S_6} \end{bmatrix}$$

$$\begin{bmatrix} \mu^{S_1} \\ \mu^{S_2} \\ \mu^{S_3} \\ \mu^{S_4} \\ \mu^{S_5} \\ \mu^{S_6} \end{bmatrix} = \begin{bmatrix} \mu^{S_3} + \mu^{S_4} - \mu^{S_6} \\ \mu^{S_3} + \mu^{S_4} - \mu^{S_5} \\ \mu^{S_3} \\ \mu^{S_4} \\ \mu^{S_5} \\ \mu^{S_6} \end{bmatrix}$$

For instance, suppose that $\mu^{S_3},\ \mu^{S_4},\ \mu^{S_5}, \text{and } \mu^{S_6}$. Let $\mu^{S_3} = 5,\ \mu^{S_4} = 6,\ \mu^{S_5} = 7, \text{and } \mu^{S_6} = 8$. Then

$$\begin{aligned} &\implies \mu^{S_1} = \mu^{S_3} + \mu^{S_4} - \mu^{S_6} = 5 + 6 - 8 = 3 \\ &\implies \mu^{S_2} = \mu^{S_3} + \mu^{S_4} - \mu^{S_5} = 5 + 6 + -7 = 4 \end{aligned}$$

Thus,

$$\begin{aligned} \mu_1 &= n_1^{-1} \sum_{j=1}^{n_1} \mu_{1j} = \frac{3+4+5+6+7+8}{6} = 5.5 \\ \mu_2 &= n_2^{-1} \sum_{j=1}^{n_2} \mu_{2j} = \frac{3(5)+3(6)}{6} = 5.5 \\ \mu_3 &= n_3^{-1} \sum_{j=1}^{n_3} \mu_{3j} = \frac{3(3)+3(8)}{6} = 5.5 \end{aligned}$$

as required by the two-level the mean condition.

The final step in mapping the vector of free set means $\boldsymbol{\mu}_0$, and by extension the vector of set means $\boldsymbol{\mu}_{\text{var}}$, to the vector of observations and thus to the $\boldsymbol{\mu}$ vector of length $\sum_{i=1}^{g} n_i$, is to identify how the individual-level set means are mapped to each μ_{ij}. By construction, $\mathbf{G}$ accomplishes this goal:

$$\Big[\mu_{11}, \mu_{12}, ..., \mu_{1n_1}, \mu_{21}, ..., \mu_{gn_g}\Big]^T = \boldsymbol{\mu} = \mathbf{G}^T \boldsymbol{\mu}_{\text{var}}$$

Substituting for μ_{var} yields:

$$\begin{aligned}
\boldsymbol{\mu} &= \mathbf{G}^T \boldsymbol{\mu}_{\text{var}} \\
&= \mathbf{G}^T \Big[\mathbf{D}\Big(\mathbf{AD}\Big)^{-1}\Big(- \mathbf{AE}\boldsymbol{\mu}_0\Big) + \mathbf{E}\boldsymbol{\mu}_0 \Big] \\
&= \mathbf{G}^T \Big[- \mathbf{D}\Big(\mathbf{AD}\Big)^{-1}\mathbf{AE} + \mathbf{E} \Big] \boldsymbol{\mu}_0 \\
&= \mathbf{H}\boldsymbol{\mu}_0
\end{aligned}$$

here illustrated by

$$\boldsymbol{\mu} = \begin{bmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 1 & 1 & 1
\end{bmatrix}^T \begin{bmatrix} \mu^{S_3} \\ \mu^{S_4} \\ \mu^{S_5} \\ \mu^{S_6} \end{bmatrix}$$

Accordingly, for any pair of partitions, a corresponding $\mathbf{H}$ may be specified. This term is utilized not just in defining $P(\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y})$ and $P(\rho_{\text{g}}, \rho_{\text{ind}}|\mathbf{Y})$, but also in the estimation of the observation means which will be detailed in the next chapter.

5 Computational Approach to Estimating the Observation Means

As stated above, two different estimates of the observation means are provided: the two-level product estimate and the posterior mode. The former provides a weighted average of posterior means over all possible partition pairs, and the latter provides the posterior mean for the pair of partitions with the maximum *a posteriori* density. Contextual reasons may provide rationale for choosing one over the other.

Given a pair of partitions, a posterior mean may be calculated from $E(\boldsymbol{\mu}_0|\rho_{\mathrm{g}}, \rho_{\mathrm{ind}}, \mathbf{Y})$ and the partitions' corresponding $\mathbf{H}$ matrix for every item. If accurate, these values would exactly satisfy the two-level mean condition. The two-level product estimate is a weighted average over all possible partition pairs of these conforming mean estimates. By construction, the two-level product estimate incorporates uncertainty in the clustering of the items and group-level indices into the estimate, but does not result in any specific partition pair in conjuction with the estimate.

In contrast, the posterior mode provides estimates of the item means which would themselves conform to the two-level mean condition **and** a specific partitioning of the items and group-level indices. However, this estimate does not explicitly incorporate

alternative partition pairs and their associated mean estimates.

In the one-level setting, the product estimate has been termed the "gold standard" ([QI03]) of estimates against which other estimates may be compared and has a prominant place in the literature. As such, emphasis will be given to examining the two-level product estimate in the two-level setting.

From settings $\mathcal{A}$ (known σ^2) and $\mathcal{B}$ (unknown σ^2) in Chapter 3, $E(\boldsymbol{\mu}_0|\rho_{\mathrm{g}}, \rho_{\mathrm{ind}}, \mathbf{Y})$ is fully specified, as is $\mathbf{H}$ per the previous chapter. Since $\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\mu}_0$, $\boldsymbol{\mu}$ is estimated with $\hat{\boldsymbol{\mu}}$, which is termed herein the *two-level product estimate*, where

$$\hat{\boldsymbol{\mu}} = \sum \mathbf{H} E(\boldsymbol{\mu}_0|\rho_{\mathrm{g}}, \rho_{\mathrm{ind}}, \mathbf{Y}) P(\rho_{\mathrm{g}}, \rho_{\mathrm{ind}}|\mathbf{Y}) \tag{5.1}$$

where the summation is over all possible pairs of partitions.

For even moderate numbers of items, it is impractical to sum over all possible pairs of partitions. For any set of n items, the number of partitions, $\mathrm{r}_{n,\gamma}$, of those items into γ sets is a Stirling Number of the Second Kind:

$$\mathrm{r}_{n,\gamma} = \frac{1}{\gamma!} \sum_{j=0}^{\gamma} (-1)^j \binom{\gamma}{j} (\gamma - j)^n$$

As n increases, $\sum_{\gamma=1}^{n} \mathrm{r}_{n,\gamma}$ (the n^{th} Bell number) increases at a rate faster than exponential ([Cro97]), and the enumeration of the total number of partitions for the calculation of a one-level product estimate becomes increasingly impractical. This is the case to an even greater degree in the two-level setting as the total number of partitions is

$$\mathrm{r}_{g,\phi,q,\gamma} = \left(\sum_{\phi=1}^{g} \mathrm{r}_{g,\phi} \right) \left(\sum_{\gamma=1}^{q} \mathrm{r}_{q,\gamma} \right)$$

In [Cro97], the one-level product estimate is estimated through Markov sampling

which is also a suitable strategy in the two-level setting. Accordingly, Markov sampling is used to approximate the two-level product estimate (5.1) according to the sampling procedure detailed below.

Markov Sampling

Each full step in the Markov chain constitutes a series of substeps motivated by and similar to those contained in [Cro97]. Like [Cro97], each substep considers partition transitions. In the two-level setting, however, transitions occur from a pair of partitions, ρ_{g} and ρ_{ind}, to another pair, and they both transition simultaneously utilizing $P(\rho_{\mathrm{g}}, \rho_{\mathrm{ind}}|\mathbf{Y})$ to govern transition probabilities. Once a full step is completed, the current partition is noted, and $\mathbf{H}E(\boldsymbol{\mu}_0|\rho_{\mathrm{g}}, \rho_{\mathrm{ind}}, \mathbf{Y})$ is calculated (recall $\mathbf{H}$ changes as a function of ρ_{g} and ρ_{ind}). For example, if at the end of the η^{th} step, the current partition pair is $\rho_{\mathrm{g}}^{\eta}, \rho_{\mathrm{ind}}^{\eta}$, then $\mathbf{H}E(\boldsymbol{\mu}_0|\rho_{\mathrm{g}}^{\eta}, \rho_{\mathrm{ind}}^{\eta}, \mathbf{Y})$ is calculated. After a sufficiently large number of steps (N), $E(\boldsymbol{\mu}|\mathbf{Y})$ is estimated by $\sum_{\eta=1}^{\mathrm{N}} \mathbf{H}E(\boldsymbol{\mu}_0|\rho_{\mathrm{g}}^{\eta}, \rho_{\mathrm{ind}}^{\eta}, \mathbf{Y})/\mathrm{N}$.

The Markov sampling procedure begins by randomly ordering the individual-level items and the group-level items (indices) separately. This ordering remains throughout the Markov sampling procedure. Begin with some pair of partitions $\rho' = \{\rho'_{\mathrm{g}}, \rho'_{\mathrm{ind}}\}$. Each full step in the sampling procedure begins by selecting the first individual-level item and the first group-level item (as ordered) to move their respective set memberships. A move for an item constitutes remaining in its current set, moving to an alternative existing set, or becoming the sole member of a new set. The possible moves of these items define the possible new partition pairs to which ρ'_{g} and ρ'_{ind} may transition in one substep. The transition probability from ρ' to one of the possible new pairs $\rho^* = \{\rho^*_{\mathrm{g}}, \rho^*_{\mathrm{ind}}\}$ is proportional to $P(\rho^*_{\mathrm{g}}, \rho^*_{\mathrm{ind}}|\mathbf{Y})$, normalized by the sum of the posterior densities of the possible new pairs. The next substep transitions from ρ^* to a new pair of partitions by moving the next group

and individual-level items as determined by the random ordering. Let *lcm*, be the least common multiple of the number of individual-level items, $q = \sum_{i=1}^{g} n_i$, and the number of groups, g. For purposes of homogeneity explained below, a full step is complete when *lcm* substeps have occurred, looping through the items at each level in order as necessary. Note that if the data are balanced with $n_i = n \ \forall i$, then a full step occurs after all individual-level items have been moved once. Once a full step is completed, the current partition is noted, and $\mathbf{H}E(\boldsymbol{\mu}_0|\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y})$ is calculated. The next step begins with the most recent pair of partitions and again moves the first pair of the ordered items. The following provides the details for the possible movements of the u^{th} individual-level item. The u^{th} group-level item (index) moves in a similar fashion. A helpful example follows the general presentation.

Movement of the u^{th} Item

Let the current individual-level partition be $\rho'_{\text{ind}} = \{S_1, ..., S_k\}$ and let the u^{th} item be in the j^{th} set, S_j. Let n_{S_j} be the number of items in set S_j. A new partition ρ^*_{ind} is formed by moving the u^{th} item. One of four cases, Case Ia, Case Ib, Case IIa, or Case IIb detailed below, holds.

To establish each of the four cases, first remove the u^{th} item to form a collection of sets $\{S^*_1, S^*_2, ..., S^*_b\}$:

Case I :

$n_{s_j} > 1 : \ S^*_r = S_r$ where r $= 1, ..., j-1, j+1, ..., k$

$S^*_j = S_j - \{u\}, \ b = k$

Case II :

$n_{s_j} = 1 : \ S^*_r = S_r$ where r $= 1, ..., j-1, j+1, ..., k-1$

$$S_j^* = S_k,\ b = k-1$$

$$\implies \{S_1^*, ..., S_{j-1}^*, S_j^*, S_{j+1}^* ..., S_b^*, \cancel{S_{b+1}}\} = \{S_1, ..., S_{j-1}, S_k, S_{j+1} ..., S_{k-1}\}$$

Next form the new partition ρ_{ind}^* which includes the u^{th} item by adding the u^{th} item to the collection of sets $\{S_1^*, ..., S_b^*\}$ in one of the two following ways:

a :

$$\rho_{\text{ind}}^* = \{S_1^*, ..., \{S_t^*, u\}, ..., S_b^*\},\ t = 1, ..., b$$

b :

$$\rho_{\text{ind}}^* = \{S_1^*, ..., S_b^*, S_{b+1}^*\},\ \text{where } S_{b+1}^* = \{u\}$$

Together, these constitute the four possible cases:

Case Ia :

$$\rho_{\text{ind}}^* = \{S_1^*, ..., \{S_t^*, u\}, ..., S_j^* = S_j - \{u\}, ..., S_b^*\},\ t = 1, ..., b$$

$$\implies = \{S_1^*, ..., \{S_t^*, u\}, ..., S_j^* = S_j - \{u\}, ..., S_k^*\},\ t = 1, ..., k,$$

$$S_r^* = S_r,\ r = 1, ..., t-1, t+1, ..., j-1, j+1, ..., k,\ S_j^* = S_j \text{ iff t=j}$$

$$\implies \textbf{Total number of sets remains the same}$$

Case Ib :

$$\rho_{\text{ind}}^* = \{S_1^*, ..., S_j^* = S_j - \{u\}, ..., S_b^*, S_{b+1}^*\},$$

$$\implies = \{S_1^*, ..., S_j^* = S_j - \{u\}, ..., S_k^*, S_{k+1}^*\},$$

$$S_r^* = S_r,\ r = 1, ..., j-1, j+1, ..., k,\ S_{k+1}^* = \{u\}$$

$$\implies \textbf{Total number of sets increases by one}$$

Case IIa :

$$\rho^*_{\text{ind}} = \{S^*_1, ..., \{S^*_t, u\}, ..., S^*_j = S_k, ..., S^*_b\},$$

$$\Longrightarrow = \{S^*_1, ..., \{S^*_t, u\}, ..., S^*_j = S_k, ..., S^*_{k-1}\},\ t = 1, ..., k-1$$

$$S^*_r = S_r,\ r = 1, ..., t-1, t+1, ..., j-1, j+1, ..., k-1,$$

$$S^*_j = S_k \text{ unless } t = j \text{ in which case } S^*_j = \{S_k, u\}$$

$$\Longrightarrow \textbf{ Total number of sets decreases by one}$$

Case IIb :

$$\rho^*_{\text{ind}} = \{S^*_1, ..., S^*_j, ..., S^*_b, S^*_{b+1}\},$$

$$\Longrightarrow = \{S^*_1, ..., S^*_j = S_k, ..., S^*_{k-1}, S^*_k\},$$

$$S^*_r = S_r,\ r = 1, ..., j-1, j+1, ..., k-1,\ S^*_j = S_k,\ S^*_k = \{u\}$$

$$\Longrightarrow \textbf{ Total number of sets remains the same}$$

The possible moves for each group-level item (index) are similar. The new partition pair will be equal to $\rho^* = \{\rho^*_{\text{g}}, \rho^*_{\text{ind}}\}$ with probability proportional to $P(\rho_{\text{g}} = \rho^*_{\text{g}}, \rho_{\text{ind}} = \rho^*_{\text{ind}}|\mathbf{Y})$. The constant of proportionality is chosen so that the sum of the transition probabilities equals 1. Each of the transition partition pairs is of one of four types:

Type 1 :

$$Q^1(t,s) = P(\rho_{\text{g}} = \{T^*_1, ..., \{T^*_t, u^{\text{g}}\}, ..., T^*_{b_{\text{g}}}\},$$
$$\rho_{\text{ind}} = \{S^*_1, ..., \{S^*_s, u^{\text{ind}}\}, ...S^*_{b_{\text{ind}}}\}|\mathbf{Y}),$$
$$t = 1, ..., b_{\text{g}},\ s = 1, ..., b_{\text{ind}}$$

Type 2 :

$$Q^2(b_{\mathrm{g}}+1,s) = P(\rho_{\mathrm{g}} = \{T_1^*,...,T_{b_{\mathrm{g}}}^*, T_{b_{\mathrm{g}}+1}^* = \{u^{\mathrm{g}}\}\},$$
$$\rho_{\mathrm{ind}} = \{S_1^*,...,\{S_s^*, u^{\mathrm{ind}}\},...S_{b_{\mathrm{ind}}}^*\}|\mathbf{Y}),$$
$$s = 1,...,b_{\mathrm{ind}}$$

Type 3 :

$$Q^3(t,b_{\mathrm{ind}}+1) = P(\rho_{\mathrm{g}} = \{T_1^*,...,\{T_t^*, u^{\mathrm{g}}\},...,T_{b_{\mathrm{g}}}^*\},$$
$$\rho_{\mathrm{ind}} = \{S_1^*,...,S_{b_{\mathrm{ind}}}^*, S_{b_{\mathrm{ind}}+1}^* = \{u^{\mathrm{ind}}\}\}|\mathbf{Y}),$$
$$t = 1,...,b_{\mathrm{g}}$$

Type 4 :

$$Q^4(b_{\mathrm{g}}+1,b_{\mathrm{ind}}+1) = P(\rho_{\mathrm{g}} = \{T_1^*,...,T_{b_{\mathrm{g}}}^*, T_{b_{\mathrm{g}}+1}^* = \{u^{\mathrm{g}}\}\},$$
$$\rho_{\mathrm{ind}} = \{S_1^*,...,,S_{b_{\mathrm{ind}}}^*, S_{b_{\mathrm{ind}}+1}^* = \{u^{\mathrm{ind}}\}\}|\mathbf{Y}),$$

Let Q^{sum} equal the sum of transition probabilities for the possible transition partition pairs.

Then the transition probability of $\rho' = \{\rho'_{\mathrm{g}}, \rho'_{\mathrm{ind}}\}$ to $\rho^* = \{\rho^*_{\mathrm{g}}, \rho^*_{\mathrm{ind}}\} =$

1. If ρ^* is of Type 1, $\frac{Q^1(t,s)}{Q^{sum}}$

2. If ρ^* is of Type 2, $\frac{Q^2(b_{\mathrm{g}}+1,s)}{Q^{sum}}$

3. If ρ^* is of Type 3, $\frac{Q^3(t,b_{\mathrm{ind}}+1)}{Q^{sum}}$

4. If ρ^* is of Type 4, $\frac{Q^4(b_{\mathrm{g}}+1,b_{\mathrm{ind}}+1)}{Q^{sum}}$

A partition pair then may be selected by sampling the possible partition pairs with weights equal to the transition probabilities.

An Example

The context of an illustrative example for the substep moving the u_{g}^{th} and u_{ind}^{th} items is helpful.

Let the current partition pair, $\rho' = \{\rho'_{\mathrm{g}}, \rho'_{\mathrm{ind}}\}$ be

$$\rho'_{\mathrm{g}} = \{1, 2, 3\}$$

$$\rho'_{\mathrm{ind}} = \Big\{\{11, 31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36\}\Big\}$$

alternatively written:

$$\rho'_{\mathrm{g}} = \{T_1\},\ T_1 = \{1, 2, 3\}$$

$$\rho'_{\mathrm{ind}} = \Big\{S_1, S_2, ..., S_6\Big\},$$

$$S_1 = \{11, 31, 32, 33\},\ S_2 = \{12\},\ S_3 = \{13, 21, 22, 23\},$$

$$S_4 = \{14, 24, 25, 26\},\ S_5 = \{15\},\ S_6 = \{16, 34, 35, 36\}$$

If $u_{\mathrm{ind}} = 11$, then $j_{\mathrm{ind}} = 1$ and $n_{S_1} > 1$, and thus Case I with $k_{\mathrm{ind}} = 6$ such that

$$S_r^* = S_r \text{ where } r = 2, 3, ..., 6$$

$$S_1^* = S_1 - \{11\},\ b = 6,$$

So the possible new individual-level partitions are:

Ia:

$$\Big\{\{\textcircled{11}, 31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36\}\Big\}$$
$$\Big\{\{31, 32, 33\}, \{12, \textcircled{11}\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36\}\Big\}$$
$$\Big\{\{31, 32, 33\}, \{12\}, \{13, 21, 22, 23, \textcircled{11}\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36\}\Big\}$$
$$\Big\{\{31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26, \textcircled{11}\}, \{15\}, \{16, 34, 35, 36\}\Big\}$$
$$\Big\{\{31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15, \textcircled{11}\}, \{16, 34, 35, 36\}\Big\}$$
$$\Big\{\{31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36, \textcircled{11}\}\Big\}$$

Ib:

$$\Big\{\{31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36\}, \{\textcircled{11}\}\Big\}$$

At the group-level, consider the possible moves the first group label may make. Denote this chosen label as $u_{\mathrm{g}} = 1$. Then $j_{\mathrm{g}} = 1$ and $n_{T_1} > 1$, and thus in Case I with $k_{\mathrm{g}} = 1$ such that

$$T_1^* = T_1 - \{1\},\ b_g = 1$$

So the possible new group-level partitions are: Ia: $\Big\{\{\textcircled{1}, 2, 3\}\Big\}$, and Ib: $\Big\{\{2, 3\}, \{\textcircled{1}\}\Big\}$. Together, the fourteen possible transition partition pairs are:

(1) $\rho_{\mathrm{g}} = \Big\{\{\textcircled{1}, 2, 3\}\Big\}$,

$$\rho_{\text{ind}} = \Big\{\{\textcircled{11}, 31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36\}\Big\}$$

(2) $\rho_{\text{g}} = \Big\{\{\textcircled{1}, 2, 3\}\Big\}$,

$$\rho_{\text{ind}} = \Big\{\{31, 32, 33\}, \{12, \textcircled{11}\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36\}\Big\}$$

(3) $\rho_{\text{g}} = \Big\{\{\textcircled{1}, 2, 3\}\Big\}$,

$$\rho_{\text{ind}} = \Big\{\{31, 32, 33\}, \{12\}, \{13, 21, 22, 23, \textcircled{11}\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36\}\Big\}$$

(4) $\rho_{\text{g}} = \Big\{\{\textcircled{1}, 2, 3\}\Big\}$,

$$\rho_{\text{ind}} = \Big\{\{31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26, \textcircled{11}\}, \{15\}, \{16, 34, 35, 36\}\Big\}$$

(5) $\rho_{\text{g}} = \Big\{\{\textcircled{1}, 2, 3\}\Big\}$,

$$\rho_{\text{ind}} = \Big\{\{31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15, \textcircled{11}\}, \{16, 34, 35, 36\}\Big\}$$

(6) $\rho_{\text{g}} = \Big\{\{\textcircled{1}, 2, 3\}\Big\}$,

$$\rho_{\text{ind}} = \Big\{\{31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36, \textcircled{11}\}\Big\}$$

(7) $\rho_{\mathrm{g}} = \Big\{\{2,3\},\{\textcircled{1}\}\Big\}$,

$\rho_{\mathrm{ind}} = \Big\{\{\textcircled{11},31,32,33\},\{12\},\{13,21,22,23\},$

$\{14,24,25,26\},\{15\},\{16,34,35,36\}\Big\}$

(8) $\rho_{\mathrm{g}} = \Big\{\{2,3\},\{\textcircled{1}\}\Big\}$,

$\rho_{\mathrm{ind}} = \Big\{\{31,32,33\},\{12,\textcircled{11}\},\{13,21,22,23\},$

$\{14,24,25,26\},\{15\},\{16,34,35,36\}\Big\}$

(9) $\rho_{\mathrm{g}} = \Big\{\{2,3\},\{\textcircled{1}\}\Big\}$,

$\rho_{\mathrm{ind}} = \Big\{\{31,32,33\},\{12\},\{13,21,22,23,\textcircled{11}\},$

$\{14,24,25,26\},\{15\},\{16,34,35,36\}\Big\}$

(10) $\rho_{\mathrm{g}} = \Big\{\{2,3\},\{\textcircled{1}\}\Big\}$,

$\rho_{\mathrm{ind}} = \Big\{\{31,32,33\},\{12\},\{13,21,22,23\},$

$\{14,24,25,26,\textcircled{11}\},\{15\},\{16,34,35,36\}\Big\}$

(11) $\rho_{\mathrm{g}} = \Big\{\{2,3\},\{\textcircled{1}\}\Big\}$,

$\rho_{\mathrm{ind}} = \Big\{\{31,32,33\},\{12\},\{13,21,22,23\},$

$\{14,24,25,26\},\{15,\textcircled{11}\},\{16,34,35,36\}\Big\}$

(12) $\rho_{\mathrm{g}} = \Big\{\{2,3\},\{\textcircled{1}\}\Big\}$,

$\rho_{\mathrm{ind}} = \Big\{\{31,32,33\},\{12\},\{13,21,22,23\},$

$\{14,24,25,26\},\{15\},\{16,34,35,36,\textcircled{11}\}\Big\}$

$$(13)\ \rho_{\mathrm{g}} = \Big\{\{\textcircled{1}, 2, 3\}\Big\},$$
$$\rho_{\mathrm{ind}} = \Big\{\{31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36\}, \{\textcircled{11}\}\Big\}$$

$$(14)\ \rho_{\mathrm{g}} = \Big\{\{2, 3\}, \{\textcircled{1}\}\Big\},$$
$$\rho_{\mathrm{ind}} = \Big\{\{31, 32, 33\}, \{12\}, \{13, 21, 22, 23\}, \{14, 24, 25, 26\}, \{15\}, \{16, 34, 35, 36\}, \{\textcircled{11}\}\Big\}\Big\}$$

Using the posterior distribution for the pair of partitions, $P(\rho_{\mathrm{g}}, \rho_{\mathrm{ind}}|\mathbf{Y})$, the transition probabilities for each possible transition partition pair may be calculated (each normalized by the sum of the fourteen respective density values, Q^{sum}) . Using these probabilities as weights, the new partition pair $\rho^* = \{\rho^*_{\mathrm{g}}, \rho^*_{\mathrm{ind}}\}$ may be randomly selected.

These substeps proceed, moving the next individual-level item and group index, until *lcm* substeps have been completed, constituting the end of the step. At this point, $\mathbf{H}E(\boldsymbol{\mu}_0|\rho_{\mathrm{g}}, \rho_{\mathrm{ind}}|\mathbf{Y})$ may be calculated using the partition pair selected at the end of the step. The next step begins by again considering the possible moves of the first pair of individual and group-level items.

The Limiting Distribution of the Markov Chain

The following establishes several properties of the Markov chain and its distribution. Together these properties imply a unique and stationary limiting distribution of the Markov chain determined by the partition pair transitions and their associated probabilities.

The transition probability associated with transitioning from the pair of partitions ρ' to ρ^* may alternatively be expressed with probability

$$\frac{\pi(\rho^* = \{\rho^*_{\text{g}}, \rho^*_{\text{ind}}\}|\mathbf{Y})}{\sum_{\rho_0|\rho_0 \in R_{\rho'}} \pi(\rho_0|\mathbf{Y})} \tag{5.2}$$

where $R^{\rho'}_{u_{\text{g}},u_{\text{ind}}}$ is the set of pairs of partitions which can be obtained from ρ' after moving the u^{th}_{g} and the u^{th}_{ind} items in their respective partitions. Following the stationarity proof in [Cro97], after moving the u^{th}_{g} and the u^{th}_{ind} items and transitioning from ρ' to ρ^*, let ρ^* have probability distribution π_1.

$$\begin{aligned}
&\pi_1(\rho^* = (\rho^*_{\text{g}}, \rho^*_{\text{ind}})|\mathbf{Y}) \\
&\quad = \sum_{\rho'} \pi_1(\rho^* = \{\rho^*_{\text{g}}, \rho^*_{\text{ind}}\}|\rho' = \{\rho'_{\text{g}}, \rho'_{\text{ind}}\}, \mathbf{Y})\pi(\rho'|\mathbf{Y}) \\
&\quad = \sum_{\rho' \in R^{\rho'}_{u_{\text{g}},u_{\text{ind}}}} \pi_1(\rho^*|\rho', \mathbf{Y})\pi(\rho'|\mathbf{Y}), \\
&\qquad \text{since if } \rho' \notin R^{\rho^*}_{u_{\text{g}},u_{\text{ind}}}, \text{ then } \pi_1(\rho^*|\rho', \mathbf{Y}) = 0 \\
&\quad = \sum_{\rho' \in R^{\rho^*}_{u_{\text{g}},u_{\text{ind}}}} \frac{\pi(\rho^*|\mathbf{Y})}{\sum_{\rho_0|\rho_0 \in R_{\rho'}} \pi(\rho_0|\mathbf{Y})}\pi(\rho'|\mathbf{Y}), \\
&\qquad \text{from definition of transition probabilities (5.2)} \\
&\quad = \sum_{\rho' \in R^{\rho^*}_{u_{\text{g}},u_{\text{ind}}}} \frac{\pi(\rho^*|\mathbf{Y})}{\sum_{\rho_0|\rho_0 \in R^{\rho^*}_{u_{\text{g}},u_{\text{ind}}}} \pi(\rho_0|\mathbf{Y})}\pi(\rho'|\mathbf{Y}), \\
&\qquad \text{since } \rho' \in R^{\rho^*}_{u_{\text{g}},u_{\text{ind}}} \implies R^{\rho'}_{u_{\text{g}},u_{\text{ind}}} = R^{\rho^*}_{u_{\text{g}},u_{\text{ind}}} \\
&\quad = \pi(\rho^*|\mathbf{Y}) \frac{\sum_{\rho' \in R^{\rho^*}_{u_{\text{g}},u_{\text{ind}}}} \pi(\rho'|\mathbf{Y})}{\sum_{\rho_0|\rho_0 \in R^{\rho^*}_{u_{\text{g}},u_{\text{ind}}}} \pi(\rho_0|\mathbf{Y})}
\end{aligned}$$

$$= \pi(\rho^*|\mathbf{Y})$$

Thus, the resulting distribution after moving a pair of items is always π. Therefore, for a full step in the Markov chain in which successively *lcm* pairs of items are moved, the resulting distribution for the pair of partitions at the end of that step is also π.

QED

It is important that there are *lcm* ordered substeps[1] in each full step, especially when the group sizes are unbalanced. For any pair of partitions, the possible transition partitions resulting from one substep depend on the group and individual-level items being moved in that substep. To illustrate, in the example above, the possible transition partitions would be different if some other pair of items were being moved rather than the first pair. If the increment of the chain were instead the substeps, the chain would not necessarily be time homogenous such that for some arbitrary number of increments v,

$$P(\rho^{\eta+v} = \{\rho^*_{\mathrm{g}}, \rho_{\mathrm{ind}}\}^*|\rho^{\eta} = \{\rho'_{\mathrm{g}}, \rho'_{\mathrm{ind}}\}) = P(\rho^{v} = \{\rho^*_{\mathrm{g}}, \rho^*_{\mathrm{ind}}\}|\rho^{0} = \{\rho'_{\mathrm{g}}, \rho'_{\mathrm{ind}}\})$$

Further, in the two-level setting it is not sufficient to make the general rule that a full step be complete after all of the individual-level items have been moved once as in [Cro97] for the one-level setting. This sufficices if the group sizes are balanced with $n_i = n\ \forall i$, but breaks down for imbalanced data. In this latter case, the problem resides in which initial group-level item is to be moved in each step. Beginning with the next group-level item after the one moved in the last substep of the previous step results in a different set of paired items to be moved together in each step. Therefore, homogeneity does not hold. Beginning with the same individal and group-level

[1]The ordering is established by some initial ordering of the items which remains constant throughout the sampling procedure

items in each step ensures the same item pairings in each step and homogeneity, but early-ordered group-level items will be moved systematically more frequently than late-ordered items. Beginning each step moving the same pair of group and individual-level items and conducting *lcm* substeps ensures the same set of moved item pairs in each step, homogeneity, and balance in the number of moves for each item at their respective levels per step.

By [HPS72] Theorem 7, page 73, if a Markov chain is aperiodic, irreducible and positive recurrent with a stationary distribution, then the limiting distribution of the Markov chain is that stationary distribution. Each aspect is addressed as follows to establish that the unique stationary distribution of the above Markov chain is its limiting distribution. By [HPS72] page 73, a sufficient condition to establish that the Markov chain is aperiodic is that there is a positive probability that a partition pair can transition to itself over a full step in the chain. This is immediately satisfied. Both the number of possible individual-level partitions and the number of possible group-level partitions are finite, and thus the number of possible partition pairs is also finite. By the transition procedure above, every pair of partitions is accessible from every other pair of partitions in some finite number of transitions. As such, all the partition pairs communicate with one another, and thus the set of possible partition pairs is irreducible by definition.[2] A state ζ in a Markov chain is said to be positive recurrent if

$$\lim_{\eta \to \infty} \frac{\eta}{N_\eta(\zeta)} < \infty$$

where N_η is the number of times the Markov chain is in state ζ at times $1, ..., \eta$. By [HPS72] Theorem 3 and Corollary 2, page 62, since the set of all possible partition

[2]It is immediate that no "alternative partition states" outside the set of possible partition pairs may be reached from any of the possible partition pairs. As such, the possible partition pairs constitute a closed set of states.

pairs is finite and irreducible, the Markov chain resulting from the above procedure is a positive recurrent chain. Finally, [HPS72] Corollary 5, page 67, since the Markov chain is irreducible and has a finite number of states, it has a unique stationary distribution which from above is π.

Posterior Mode

Through the course of the Markov sampling procedure, the joint posterior density of many pairs of partitions is calculated. Keeping track of which pair yields the largest evaluated density is straightforward and inexpensive. At the end of the Markov sampling procedure, this partition pair, $\rho_{\mathrm{g}}^{\mathrm{MAP}}, \rho_{\mathrm{ind}}^{\mathrm{MAP}}$ may be used as an estimate of the maximum *a posteriori* partition pair, and $\boldsymbol{\mu}$ may be estimated by $\mathbf{H}E(\boldsymbol{\mu}_0|\rho_{\mathrm{g}}^{\mathrm{MAP}}, \rho_{\mathrm{ind}}^{\mathrm{MAP}}, \mathbf{Y})$. This is resonably costless when done in conjunction with computing the two-level product estimate.

In [QI03], there is a concern stemming from the fact that with a large number of partitions the posterior probability for each possible partition is going to be extremely small. This notion extends to the two-level setting. As such, when model complexity is a concern and the desired optimum solution is one that minimizes some loss function which balances the largest posterior density against the number of clusters, attempting to find the general posterior mode may not be the best option. [QI03] suggests a way to address this issue via a two-stage algorithm which first estimates a product estimate for item observation means and then attempts to identify a partition which minimizes a loss function weighing a comparision between the product estimate and the partition specific mean estimates against model complexity. As computational concerns in the two-level setting grow, a way to avoid the costly "two-stage" implementation would be to consider partition specific information while conducting the Markov sampling as noted in the preceeding paragraph. A

potential strategy is to choose the partition pair from among the N partition pairs used to estimate the product estimate. The chosen partition pair would yield mean estimates which minimize the selected loss function.

In Chapter 3, the posterior distribution for $\boldsymbol{\mu}_0$, given the data and a specified pair of partitions ρ_{g} and ρ_{ind}, is shown to be Normally distributed. In this process, both $E(\boldsymbol{\mu}_0|\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y})$ and $Cov(\boldsymbol{\mu}_0|\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y})$ are identified as well. Accordingly, a credible interval set may be established for each free set mean utilizing the respective posterior expected values. The respective variances may be used such that the $1-\alpha$ credible interval set for each free mean μ_0^S, is:

$$E(\mu_0^S|\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y}) - z_{\alpha/2}\sqrt{Var(\mu_0^S|\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y})} \leq \mu_0^S \leq E(\mu_0^S|\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y}) + z_{\alpha/2}\sqrt{Var(\mu_0^S|\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y})}$$

Through rules of means and variances, similar credible interval sets may be established for the linearly related set means $\boldsymbol{\mu}_{\text{rel}}$.

6 Simulated Data Results

Simulating data for use in conducting a two-level PPM analysis is straightforward once true underlying partitions are defined. **R** has many built in functions for simulating observations from univariate and multivariate normal distributions which simply require the specification of means and variances (or a covariance matrix) for use. If the observation variances are assumed known, these may be set immediately by the user. Given a particular pair of true group and individual-level partitions, a modeler may carefully specify generating μ_{ij}'s so that they appropriately conform to the two-level mean condition implied by the partitions. Alternatively, based on the true partitions, the modeler may go through the process of determining which individual-level partition sets and associated means are free and which are linearly related and randomly generate $\boldsymbol{\mu}_0$ accordingly. This vector of free means may then be mapped to the items to set the true μ_{ij}'s across all items via the corresponding $\mathbf{H}$ matrix determined by the true partition pair. Once these μ_{ij}'s have been specified, item observations may be generated.

To compute either an exact value or Markov sampling estimate of the two-level product estimate, the m values at the group and individual-levels (m_{gr} and m_{ind}), along with any other hyperparameters in the model need to be addressed. For example, when the observation variances are known as in setting $\mathcal{A}$, $\boldsymbol{\theta}^*$ and $\boldsymbol{\gamma_0}$ must be specified.

In the one-level setting, Crowley in [Cro92], [Cro97], and other authors using a Ewens-Pitman or similar partition prior must specify a single m value which serves a similar function as the two-level m values. Again, hyperparameters such as the means and variances for the prior distributions of set means must be addressed as well. The PPM literature does not have one consistent approach for doing this. The options are to assign these hyperparameters explicitly, set them empirically, set a hyperprior distribution, or some combination of the three. In [Cro92], a hyperprior distribution on each of these parameters (m, and the prior mean and variance for every set mean) is used. This prior for m is specified as a discrete prior with equal mass at values 1, 2, 4, 8, and 16. The values were chosen simply after experimentation. Alternatively, m is specified empirically in [Cro97]. Notably, [Cro97] estimates m by ordering the observations and counting the number of intervals between sequential observation values which are greater than 1. Calling this count $nsets$, she defines $m = 2^{nsets-2}$. For two-level clustering, a similar empirical method may be used for both m_{gr} and m_{ind}. At the group level, $nsets_{gr}$ may be calculated by counting the intervals between the group average observation values, $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$. However, neither a prior nor an empirical method is necessarily the most common route to addressing m. Frequently in the literature, m is simply set to 1 (see, for example, [QI03] and [CMG+14]). With a Ewens-Pitman prior partition distribution, smaller values of m promote larger clusters with relatively many items, hence, a smaller number of clusters. As noted in [Cro92] and [QI03], the results are quite sensitive to the value of m, and thus how it is addressed is important. In the two-level setting this issue is compounded. The *two* "m" hyperparameters admit the possibility of many possible combinations of ways to address their values. Later in this chapter, these different possibilities and their respective results are discussed.

As mentioned above, [Cro92] sets all of the hyperparameters in her model using

hyperprior distributions. These hyperprior distributions on the parameters of prior set mean distributions in [Cro92] are set equal to 1 to remove location effects, but admittedly mostly out of convenience. [Cro97] specifies the parameter values for the prior distribution of each set mean empirically from the data, grudgingly admitting that there is no "natural choice for priors" for these terms. In the two-level setting and in the opinion of this dissertation's author, the empirically driven method for setting these hyperparameters is accordingly preferred. In the two-level setting, the hyperparameters for the prior distribution of $\boldsymbol{\mu}_0$, $\boldsymbol{\theta}^* = \theta \mathbf{1}_w$ and $\boldsymbol{\gamma}_0 = \sigma_0^2 \boldsymbol{I}_w$, are thus estimated from the data as follows. For a given pair of partitions, the individual-level sets associated with the free means, $\mathbf{S}^{\boldsymbol{\mu}_0} = \{S^{\mu_0^1}, S^{\mu_0^2}, ..., S^{\mu_0^w}\}$, may be identified. For each of these sets, a sample mean, $\bar{y}^{S^{\mu_0^s}}$, is calculated based on the items mapped to each set $S^{\mu_0^s}$, $\{(ij) : k_1(ij) = S^{\mu_0^s}\}$, $(ij) \in \Omega_1$. Thus, there are $w = ||\boldsymbol{\mu}_0||$ sample means, one for each free set mean. From these, θ may be estimated by

$$\theta = \frac{1}{w} \sum_{i=1}^{w} \bar{y}^{S^{\mu_0^i}}$$

This results in a "de-weighted" average of the free set means capturing the range of free set observation values. σ_0^2 is estimated by

$$\sigma_0^2 = \frac{\sum\limits_{\substack{ij \\ k_1(ij) \in \mathbf{S}^{\boldsymbol{\mu}_0}}} \Big(y_{ij} - \theta\Big)^2}{\Big(\sum\limits_{\substack{ij \\ k_1(ij) \in \mathbf{S}^{\boldsymbol{\mu}_0}}} 1\Big) - 1}$$

This is a modified sample variance over the free set observation values using θ as the sample mean. There are of course alternative ways of estimating these terms, including using the overall sample mean, $\theta = \bar{y}$, or the sample mean of the free set observations, $\theta = \bar{y}^{S^{\mu_0}}$. However, on initial testing, superior results were derived

from the specified method. More aggressive forms of estimating $\boldsymbol{\theta}^*$ would include allowing its w components to vary, $\theta^{\mu_0^1}, \theta^{\mu_0^2}, ..., \theta^{\mu_0^w}$, with $\theta^{\mu_0^s} = \bar{y}^{S^{\mu_0^s}}$.

As this dissertation is meant not only to introduce the two-level setting, but also to extend the results of [Cro97] to the two-level setting, the simulations below focus on the known and constant variance setting $\mathcal{A}$. However, initial analysis of the unknown and constant variance setting $\mathcal{B}$ yield similarly positive results when compared against the one-level setting.

The following subsections present results from simulations. First, the Markov sampling approximation of the two-level product estimate is compared with the exact value. Next, the two-level product estimate is compared with the one-level product estimate, and an example of the posterior mode estimate is provided. Then follows a discussion of the robustness of the two-level product estimate to misspecification of the data distribution and to the constant variance assumption. Then the sensitivity of the results to changes in how m_{gr} and m_{ind} are specified is discussed.

6.1 Comparison of Exact and Approximate Two-Level Product Estimates

To test the Markov sampling estimation of the two-level product estimate presented in Chapter 5, it is compared with the results from exact computation of the two-level product estimate. Over a number of simulations, the results of the estimation were quite close to the exact values. Two examples are provided. Letting $g = 3$ and $n_1 = n_2 = n_3 = 3$, there are nine total items with three in each group, and a total of $5 \times 21,147 = 105,735$ possible partition pairs. The means of the generating distributions for the items, the simulated data values, the corresponding true partition structure of the items, and the absolute difference between the exact

and estimated two-level product estimates are provided in the following tables. In keeping with a certain amount of precedent, m_{gr} and m_{ind} are both set equal to 1. Table 6.1 presents an example with distinct separation in the underlying item means compared to observation variance. Table 6.2 illustrates that even when there is smaller mean separation relative to observation variance, and therefore cluster means are more difficult to identify, again, the exact and estimated two-level product estimate calculations are quite similar. This suggests that the Markov sampling approximation method works reasonably well.

Table 6.1: Comparison of Exact vs. Markov Approximation Two-Level Product Estimate Calculation, $\sigma^2 = 1$

Group Index	Individual Index	μ_{ij}	y	Absolute Difference Between $\hat{\mu}_{ij}$ and Markov Approx.
1	1	30	30.13386	0.04404
1	2	40	39.90950	0.00301
1	3	50	49.80236	0.00229
2	1	40	40.46237	0.00496
2	2	40	40.41625	0.00421
2	3	40	39.72541	0.00032
3	1	40	40.54724	0.00722
3	2	50	49.76597	0.00193
3	3	60	61.02215	0.02137

The exact value of the two-level product estimate for each individual-level mean and its Markov approximation are compared by taking the absolute difference between their values

6.2 Comparison of One-Level vs. Two-Level Product Estimates

The group membership indices provide additional information which serves to improve the estimation of the individual item observation means. Experimental simulation with comparison of one and two-level product estimates illustrates this point

Table 6.2: Comparison of Exact vs. Markov Approximation Two-Level Product Estimate Calculation, $\sigma^2 = 1$

Group Index	Individual Index	μ_{ij}	y	Absolute Difference Between $\hat{\mu}_{ij}$ and Markov Approx.
1	1	3	3.551080	0.000914
1	2	4	3.911063	0.001815
1	3	5	4.091670	0.003087
2	1	4	4.496084	0.002944
2	2	4	5.590253	0.017069
2	3	4	4.624190	0.011515
3	1	4	3.678996	0.006314
3	2	5	4.749579	0.000843
3	3	6	6.684084	0.003654

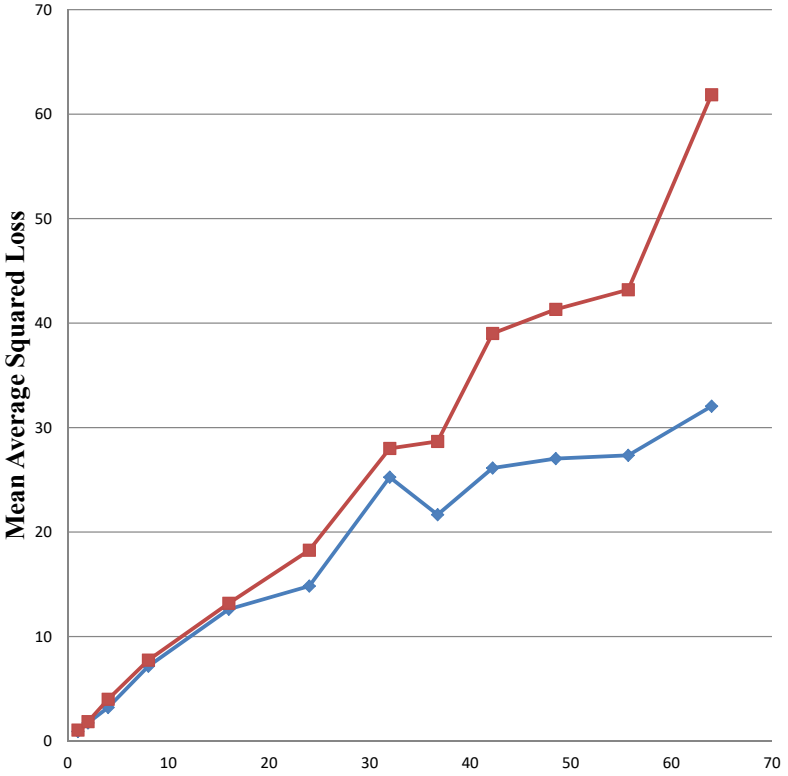
The exact value of the two-level product estimate for each individual-level mean and its Markov approximation are compared by taking the absolute difference between their values

well. A simple illustrative example follows which provides a comparison of the one-level and two-level product estimate across many different σ^2 values.

For a given partition structure, μ_{ij}'s are randomly generated such that they satisfy the two-level mean condition. Using these means, item observations are simulated using a specified constant variance across all observation distributions. One and two-level product estimates then may be calculated and the average squared loss calculated for each. For that specified constant variance σ^2, let the observation simulation be repeated for a total of ten times, and the ten average squared loss values themselves be averaged. This results in one mean average squared loss value for the one-level product estimate and one loss value for the two-level product estimate for that specific σ^2. Repeat the process, but with a different σ^2 value. Figure 6.1 provides an example of these comparisons for several different σ^2 values. As before, $m_{\text{gr}} = 1$ and $m_{\text{ind}} = 1$ with $\sigma^2 = 1$.

For greater consistency with [Cro97], it is worthwhile to estimate the individual-level "m" hyperparameter, m_{ind}, in the two-level setting empirically by the method [Cro97] uses to estimate the m hyperparameter in the one-level setting. The group-

Figure 6.1: Comparison of Mean Average Squared Loss of One-Level and Two-Level PPM Product Estimates by Observation Variance, g = 4



ij	Group Level Partition	Individual Level Partition	μ_{ij}
11	1	1	40.13205
12	1	2	37.36941
13	1	3	47.54540
14	1	4	44.78276
21	1	1	40.13205
22	1	1	40.13205
23	1	4	44.78276
24	1	4	44.78276
31	2	2	37.36941
32	2	3	47.54540
33	2	4	44.78276
34	2	5	54.95876
41	2	2	37.36941
42	2	2	37.36941
43	2	5	54.95876
44	2	5	54.95876

level "m" hyperparameter, m_{gr} remains set equal to 1, but further discussion of the effects of specifying m_{gr} alternatively will follow below. For the results that follow, the simulation is repeated 50 times. As shown in Tables 6.3 - 6.7 at the end of the chapter, specifying m_{ind} in this manner, the two-level product estimate performs admirably, providing superior (lower average squared loss) results across a variety of mean and group specifications. There is indication that the two-level product estimate performs relatively better when there are fewer groups with equal group means, μ_i, (hence larger group-level clusters) than not.

Posterior Mode Estimate

As noted above, should a specific partitioning of the items be desired as part of the result, the estimate for the posterior mode may be used. Notably, this not only provides a specific and automatic partitioning of the items into clusters, but the two-level mean condition holds explicitly in the resulting mean estimates. Figure 6.2 at the end of the chapter provides an example of a posterior mode partition with set assignments and estimates of the set means, indicating satisfaction of the two-level mean condition through these estimates. Throughout the course of experimentation, the two-level product estimate generally resulted in superior estimates over the posterior mode estimate, but this advantage tended to decrease to a point with increased sampling.

Since the means are estimated using the expected value of the posterior distribution for $\boldsymbol{\mu}_0$, this distribution may be further utilized as desired to address uncertainty through inference or construction of appropriate intervals around the estimates.

6.3 Sensitivity to Misspecification of Data Distribution

This two-level analysis makes the assumption that the distribution of each item observation is Normal. It certainly could be the case that this assumption is incorrect. As such it is warranted to check robustness relative to the misspecification of the data distribution. Interesting scenarios are to compare results from a correctly specified analysis (the data does indeed follow a Normal distribution) with results corresponding to data from a Uniform distribution and those from a t distribution. Tables 6.8 - 6.11 at the end of the chapter provide results of such comparisons for

several sets of bounds for the Uniform data and several different degrees of freedom settings for the t distribution. The correctly specified data following a Normal distribution has $\sigma^2 = 1$. When the bounds around the means of the Uniformly distributed data are $\mu_{ij} \pm 1$, the two-level product estimate of the misspecified data is superior to that of the correctly specified data. This result makes sense as the data are concentrated and restricted from being in tail sections like those of the Normal distribution. When the Uniform distribution has width $2\sqrt{3}$, and thus variance 1 so equal to that of the Normally specified data, no significant difference is found between the two-level product estimates of the misspecified and correctly specified data. Significant differences occur when the bounds around the means of the generating Uniform distribution have width 4. When the misspecified distribution is actually a t distribution, no significant difference is found between the two-level product estimates when df = 50, 30, or 20. Significant differences begin to occur when df = 10, and the number of groups with different μ_i values increases.

6.4 Sensitivity to the Constant Variance Assumption

Thus far the two-level analysis has made the assumption that item observation variance is constant across items. In otherwords, that $\sigma^2_{ij} = \sigma^2 \ \forall(ij)$. Practically, this may not be the case. However, it may be that the observation variances are similar or fall within some reasonable range. It may be desired to assume constant variance in these cases, taking a mid-point of the range of variances as a "pseudo-constant" variance. It is warranted to compare the average squared loss of the two-level product estimate when a specified variance is truly the constant variance for the items versus when that specified variance is a pseudo-constant approximation as just described. This range of variances may be estimated by many other clustering

methods, including the mixture-model approach of [FR02] and [FRS12] for example. As shown in Tables 6.12 - 6.13 at the end of the chapter, no significant differences were found in the results when a pseudo-constant variance of 1 is used when variances could range from 0.5 and 1.5, when a pseudo-constant variance of 2 is used when variances could range between 1 and 3, and when a psuedo-constant variance of 3 is used when variances could range between 1 and 5.

6.5 Sensitivity to $m_{\mathbf{gr}}$ and $m_{\mathbf{ind}}$

The results in one-level PPMs are known to vary greatly through changes in m when using cohesions of the form $c(S) = \frac{(n_S-1)!}{m^{n_s-1}}$ or in the more popular variation $c(S) = m(n_S - 1)!$. [Cro92, QI03, CMG$^+$14]

Though some authors have considered this sensitivity a weakness of this sort of prior, from an applied perspective, it affords the user a certain amount of flexibility and control over the preferred outcomes. There may be good cause to prefer a relatively small number of large clusters (partition sets) in one context versus a relatively large number of small clusters in another. In the one-level setting, deliberately adjusting m allows that particular preference a greater likelihood of being realized. For a prior with cohesions of this type, the expected number of clusters is: $E(k) = \sum_{s=1}^{n} \frac{m}{s+m-1}$ [QI03]. Figure 6.3 at the end of the chapter shows the *a priori* expected number of clusters when $n = 10, 50, 100, 500$ over several values for m. This stops short of being overly heavy handed like some popular clustering methods (ex. k-means) that require the desired number of clusters to be fully and singularly specified ahead of time.

As can be seen in Figure 6.3, larger values of m yield larger expected numbers of clusters regardless of the number of items. When there is a desired goal by the modeler, for example, to have relatively many clusters each with a small number of

items, figures like Figure 6.3 provide insight as to what an appropriate set value of m might be. This is only moreso the case in the two-level setting where preferred clustering over both group-level clusters and individual-level clusters may be influenced by changing values of m_{gr} and m_{ind}. However, if the only preference of the user is modelling some unknown true number of clusters which exist, not preferencing the outcome to either small or large clusters, and limited prior knowledge is to be had, the question remains as to how best to address m in the one-level setting in order to achieve the best estimates of the item means. Again, the issue is compounded in the two-level setting.

Care must be taken if the desire is to set m_{gr} and m_{ind} empirically. As noted, [Cro97] estimates m by counting intervals of a certain size between observations. Though guidance in [Cro97] is not provided, one quickly realizes through experimentation that in both the one and two-level settings proper calibration of this method is advisable and important. Counting the number of intervals of width X will obviously yield different values depending on X for any m hyperparameters. For many purposes, an interval width of 1 as specified in [Cro97] and used elsewhere above, is perfectly fine. However, when dealing with very small or large observation variances, this width may unintentionally obscure the partition structure by specifying far too small or large an estimate for any m hyperparameters which may exist in the model. This could unintentionally exert large pressures for few or many clusters. With a figure like Figure 6.3 in mind, considering some idea of the variance of the observations in conjunction with the interval width permits a more flexible methodology and more intentional results.

Alternatively, simply setting m in the one-level setting equal to 1 is a common occurrence for analyses in the PPM literature using Ewens-Pitman and similar partition priors especially when sensitivity to m is acknowledged (see for example [QI03] and

[CMG+14]). Aside from immediate computational advantages, m remains constant throughout experimentation regardless of the vagaries of data generated, and thus is one less moving piece to have to consider. Most importantly perhaps, setting m to be constant and small puts relative downward pressure on the total number of clusters which in turn preferences more items in fewer sets and discourages singleton clusters (sets made up of one item). This arrangement of items into fewer clusters is often a preference in the literature.

Unless otherwise noted, the simulations referenced in earlier sections have set m_{ind} empirically (using an interval size of 1) and set $m_{\text{gr}} = 1$. This was done for comparison purposes with [Cro97] at the individual item level, and is further supported by the fact that both [Cro97] and most of the simulations above set the observation variances equal to 1. At the group-level, since having the two-level setting implies the chance that some of the group indices truly belong in the same group-level partition set, m_{gr} is set to be fixed and small to discourage singleton group-level sets unless truly warranted by the evidence.

An alternative method of specifying m values comes from setting a prior on these hyperparameters. After experimentation, [Cro92] elects to specify a prior on m taking values 1, 2, 4, 8, and 16 with equal probability. This manner of addressing m hyperparameters affords some of the benefits of both fixing single constant values and setting empirically. Assigning priors of this kind allows a modeler to still preference types of partitions structures over others, but not as aggressively as specifying singular m values. Additionally, setting priors as opposed to the empirical method allows a modeler more control over the range of possible m values, notably the upper bound, regardless of the data. Accordingly, for the two-level setting, experimentation with assigning both m_{gr} and m_{ind} priors is warranted. Many different priors produced fine results over the varied mean settings tested. The most notable results

came by specifying the priors such that m_{gr} takes values 0.25, 0.50, 1, and 2 with equal probability and m_{ind} takes values 1, 2, 4, and 8 with equal probability. There is some evidence that this specification outperforms setting $m_{\text{gr}} = 1$ and m_{ind} empirically compared to the one-level product estimate, albeit in more limited testing. Indeed again, these priors generated two-level product estimates superior to those of the one-level product estimate.

Further testing of different combinations of setting priors, explicitly setting a singluar value, or empirically estimating the two m values yielded a few general themes. The results, regardless of method, were all generally similar when the m values were similar. The largest departures from one another happened when one method allowed, set, or preferenced a much higher or much lower m value than other methods. This is consistent with the one-level literature indicating general sensitivity to m. When **both** m_{gr} and m_{ind} were set empirically, results were more likely to preference partitions with far too many or too few clusters. This did not always produce worse mean estimates than other methods, but is somewhat reductive and especially visible when producing posterior mode estimates when specific group and individual-level partitions are desired as part of the results. This issue was not nearly as pronounced when m_{gr} was fixed or assigned a prior preferencing small values of m_{gr}. To that end, when using a prior, care must be taken at both levels to choose reasonable ranges of values and to weight the values in these ranges conservatively. Preferencing lower values, either through priors or fixed singular values, produced generally better results than larger values.

Another general result illustrated through experimentation is that results can vary substantially by the estimation method of m_{gr} when m_{ind} is fixed. Notably in these cases, again fixing m_{gr} to a single low value (ex. 1) or assigning a prior outperformed empirically setting m_{gr} across many observation variance values when m_{ind} was fixed

(ex. equal to 1). Together, these general results imply the importance of asserting some control over the m_{gr} value rather than allowing it unintentionally to be too large. Based on experimentation, either setting $m_{\text{gr}} = 1$ and m_{ind} empirically or specifying priors such that m_{gr} takes values 0.25, 0.50, 1, and 2 with equal probability and m_{ind} takes values 1, 2, 4, and 8 with equal probability are recommended approaches.

Table 6.3: Comparison of Mean Average Squared Loss of the One-Level and Two-Level PPM Product Estimates for a Variety of Mean and Group Specifications, $\sigma^2 = 1$

		Individual (j)					Difference Between Two-Level and One-Level Product Estimate
		1	2	3	4	Group Mean	
Group (i)	1	5	5	5	5	5	-0.020653527
	2	5	5	5	5	5	
	3	5	5	5	5	5	
	4	5	5	5	5	5	
	5	5	5	5	5	5	

		Individual (j)					Difference Between Two-Level and One-Level Product Estimate
		1	2	3	4	Group Mean	
Group (i)	1	8	10	10	12	10	-0.170640248
	2	10	10	10	10	10	
	3	8	8	12	12	10	
	4	6	8	12	14	10	
	5	6	10	10	14	10	

		Individual (j)					Difference Between Two-Level and One-Level Product Estimate
		1	2	3	4	Group Mean	
Group (i)	1	6	8	8	10	8	-0.011876583
	2	8	10	10	12	10	
	3	10	10	10	10	10	
	4	8	10	10	12	10	
	5	8	8	12	12	10	

		Individual (j)					Difference Between Two-Level and One-Level Product Estimate
		1	2	3	4	Group Mean	
Group (i)	1	6	6	10	10	8	-0.024046567
	2	6	8	8	10	8	
	3	10	10	10	10	10	
	4	8	10	10	12	10	
	5	8	8	12	12	10	

Table 6.4: Comparison of Mean Average Squared Loss of the One-Level and Two-Level PPM Product Estimates for a Variety of Mean and Group Specifications, $\sigma^2 = 1$

		Individual (j)					Difference Between Two-Level and One-Level Product Estimate
		1	2	3	4	Group Mean	
Group (i)	1	6	8	8	10	8	-0.031166636
	2	8	8	8	8	8	
	3	6	6	10	10	8	
	4	8	10	10	12	10	
	5	10	12	12	14	12	

		Individual (j)					Difference Between Two-Level and One-Level Product Estimate
		1	2	3	4	Group Mean	
Group (i)	1	6	8	8	10	8	-0.028107772
	2	6	6	10	10	8	
	3	8	10	10	12	10	
	4	8	8	12	12	10	
	5	8	10	12	14	11	

		Individual (j)					Difference Between Two-Level and One-Level Product Estimate
		1	2	3	4	Group Mean	
Group (i)	1	10	10	10	10	10	-0.019638932
	2	8	10	10	12	10	
	3	8	10	12	14	11	
	4	10	12	14	16	13	
	5	12	14	16	18	15	

		Individual (j)					Difference Between Two-Level and One-Level Product Estimate
		1	2	3	4	Group Mean	
Group (i)	1	6	8	10	12	9	-0.067642936
	2	8	10	12	14	11	
	3	10	12	14	16	13	
	4	12	14	16	18	15	
	5	14	16	18	20	17	

Table 6.5: Comparison of Mean Average Squared Loss the One-Level and Two-Level PPM Product Estimates for a Variety of Mean and Group Specifications, $\sigma^2 = 1$

		Individual (j)					Difference Between Two-Level and One-Level Product Estimate
		1	2	3	4	Group Mean	
Group (i)	1	16	20	20	24	20	-0.179881434
	2	20	20	20	20	20	
	3	16	16	24	24	20	
	4	12	16	24	28	20	
	5	12	20	20	28	20	

		Individual (j)					Difference Between Two-Level and One-Level Product Estimate
		1	2	3	4	Group Mean	
Group (i)	1	12	16	16	20	16	-0.183265723
	2	16	20	20	24	20	
	3	20	20	20	20	20	
	4	16	20	20	24	20	
	5	16	16	24	24	20	

		Individual (j)					Difference Between Two-Level and One-Level Product Estimate
		1	2	3	4	Group Mean	
Group (i)	1	12	12	20	20	16	-0.139626767
	2	12	16	16	20	16	
	3	20	20	20	20	20	
	4	16	20	20	24	20	
	5	16	16	24	24	20	

		Individual (j)					Difference Between Two-Level and One-Level Product Estimate
		1	2	3	4	Group Mean	
Group (i)	1	12	16	16	20	16	-0.041154436
	2	16	16	16	16	16	
	3	12	12	20	20	16	
	4	16	20	20	24	20	
	5	20	24	24	28	24	

Table 6.6: Comparison of Mean Average Squared Loss of the One-Level and Two-Level PPM Product Estimates for a Variety of Mean and Group Specifications, $\sigma^2 = 1$

		Individual (j)					
		1	2	3	4	Group Mean	Difference Between Two-Level and One-Level Product Estimate
Group (i)	1	12	16	16	20	16	-0.01431844
	2	12	12	20	20	16	
	3	16	20	20	24	20	
	4	16	16	24	24	20	
	5	16	20	24	28	22	

		Individual (j)					
		1	2	3	4	Group Mean	Difference Between Two-Level and One-Level Product Estimate
Group (i)	1	20	20	20	20	20	0.005033134
	2	16	20	20	24	20	
	3	16	20	24	28	22	
	4	20	24	28	32	26	
	5	24	28	32	36	30	

		Individual (j)					
		1	2	3	4	Group Mean	Difference Between Two-Level and One-Level Product Estimate
Group (i)	1	12	16	20	24	18	-0.030473034
	2	16	20	24	28	22	
	3	20	24	28	32	26	
	4	24	28	32	36	30	
	5	28	32	36	40	34	

Table 6.7: Comparison of Mean Average Squared Loss of the One-Level and Two-Level PPM Product Estimates for a Variety of Mean and Group Specifications, $\sigma^2 = 4$

		Individual (j)					
		1	2	3	4	Group Mean	Difference Between Two-Level and One-Level Product Estimate
Group (i)	1	30	40	40	50	40	-0.411464767
	2	40	50	50	60	50	
	3	50	50	50	50	50	
	4	40	50	50	60	50	
	5	40	40	60	60	50	

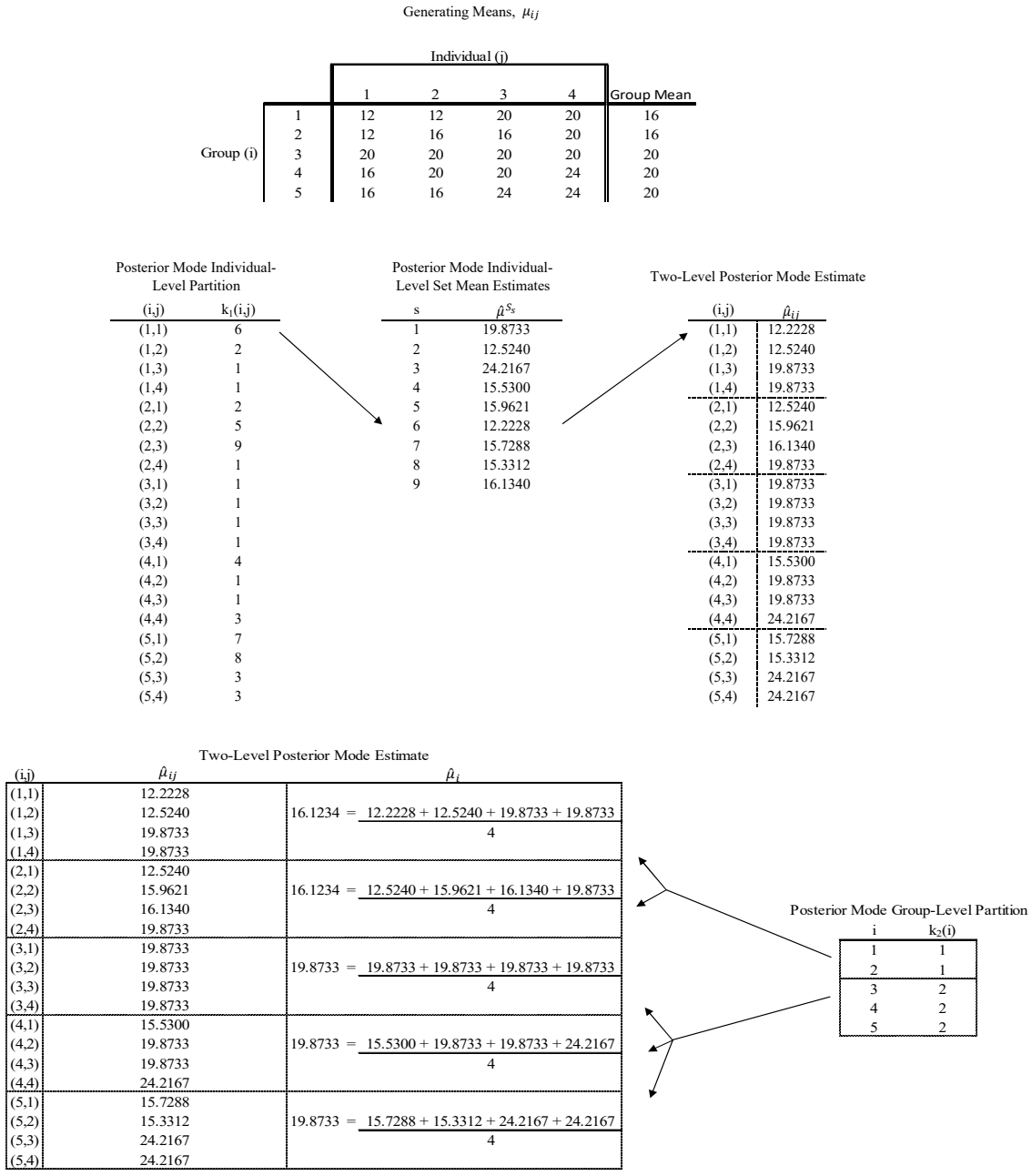
		Individual (j)					
		1	2	3	4	Group Mean	Difference Between Two-Level and One-Level Product Estimate
Group (i)	1	30	30	50	50	40	-0.376341853
	2	30	40	40	50	40	
	3	50	50	50	50	50	
	4	40	50	50	60	50	
	5	40	40	60	60	50	

		Individual (j)					
		1	2	3	4	Group Mean	Difference Between Two-Level and One-Level Product Estimate
Group (i)	1	30	40	40	50	40	-0.136714438
	2	40	40	40	40	40	
	3	30	30	50	50	40	
	4	40	50	50	60	50	
	5	50	60	60	70	60	

		Individual (j)					
		1	2	3	4	Group Mean	Difference Between Two-Level and One-Level Product Estimate
Group (i)	1	30	40	40	50	40	-0.049147566
	2	30	30	50	50	40	
	3	40	50	50	60	50	
	4	40	40	60	60	50	
	5	40	50	60	70	55	

		Individual (j)					
		1	2	3	4	Group Mean	Difference Between Two-Level and One-Level Product Estimate
Group (i)	1	50	50	50	50	50	0.04231148
	2	40	50	50	60	50	
	3	40	50	60	70	55	
	4	50	60	70	80	65	
	5	60	70	80	90	75	

Figure 6.2: Example of Two-Level Posterior Mode Estimate and Partition Structure



Group (i)	1	2	3	4	Group Mean
1	12	12	20	20	16
2	12	16	16	20	16
3	20	20	20	20	20
4	16	20	20	24	20
5	16	16	24	24	20

(i,j)	$k_1(i,j)$
(1,1)	6
(1,2)	2
(1,3)	1
(1,4)	1
(2,1)	2
(2,2)	5
(2,3)	9
(2,4)	1
(3,1)	1
(3,2)	1
(3,3)	1
(3,4)	1
(4,1)	4
(4,2)	1
(4,3)	1
(4,4)	3
(5,1)	7
(5,2)	8
(5,3)	3
(5,4)	3

s	$\hat{\mu}^{S_s}$
1	19.8733
2	12.5240
3	24.2167
4	15.5300
5	15.9621
6	12.2228
7	15.7288
8	15.3312
9	16.1340

(i,j)	$\hat{\mu}_{ij}$
(1,1)	12.2228
(1,2)	12.5240
(1,3)	19.8733
(1,4)	19.8733
(2,1)	12.5240
(2,2)	15.9621
(2,3)	16.1340
(2,4)	19.8733
(3,1)	19.8733
(3,2)	19.8733
(3,3)	19.8733
(3,4)	19.8733
(4,1)	15.5300
(4,2)	19.8733
(4,3)	19.8733
(4,4)	24.2167
(5,1)	15.7288
(5,2)	15.3312
(5,3)	24.2167
(5,4)	24.2167

(i,j)	$\hat{\mu}_{ij}$	$\hat{\mu}_i$
(1,1)	12.2228	
(1,2)	12.5240	$16.1234 = \frac{12.2228 + 12.5240 + 19.8733 + 19.8733}{4}$
(1,3)	19.8733	
(1,4)	19.8733	
(2,1)	12.5240	
(2,2)	15.9621	$16.1234 = \frac{12.5240 + 15.9621 + 16.1340 + 19.8733}{4}$
(2,3)	16.1340	
(2,4)	19.8733	
(3,1)	19.8733	
(3,2)	19.8733	$19.8733 = \frac{19.8733 + 19.8733 + 19.8733 + 19.8733}{4}$
(3,3)	19.8733	
(3,4)	19.8733	
(4,1)	15.5300	
(4,2)	19.8733	$19.8733 = \frac{15.5300 + 19.8733 + 19.8733 + 24.2167}{4}$
(4,3)	19.8733	
(4,4)	24.2167	
(5,1)	15.7288	
(5,2)	15.3312	$19.8733 = \frac{15.7288 + 15.3312 + 24.2167 + 24.2167}{4}$
(5,3)	24.2167	
(5,4)	24.2167	

i	$k_2(i)$
1	1
2	1
3	2
4	2
5	2

Table 6.8: Robustness Check to Misspecification of Data Distribution: Comparison of Mean Average Squared Loss of the Normal vs. Uniform Distribution

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	16	20	20	24	20
	2	20	20	20	20	20
	3	16	16	24	24	20
	4	12	16	24	28	20
	5	12	20	20	28	20

Uniform Distribution $[\mu_{ij}+/- 1]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{2}]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{3}]$	Uniform Distribution $[\mu_{ij}+/- 2]$
Significant Difference (+)	No Significant Difference	No Significant Difference	Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	12	16	16	20	16
	2	16	20	20	24	20
	3	20	20	20	20	20
	4	16	20	20	24	20
	5	16	16	24	24	20

Uniform Distribution $[\mu_{ij}+/- 1]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{2}]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{3}]$	Uniform Distribution $[\mu_{ij}+/- 2]$
Significant Difference (+)	No Significant Difference	No Significant Difference	Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	12	12	20	20	16
	2	12	16	16	20	16
	3	20	20	20	20	20
	4	16	20	20	24	20
	5	16	16	24	24	20

Uniform Distribution $[\mu_{ij}+/- 1]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{2}]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{3}]$	Uniform Distribution $[\mu_{ij}+/- 2]$
Significant Difference (+)	No Significant Difference	No Significant Difference	Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	12	16	16	20	16
	2	16	16	16	16	16
	3	12	12	20	20	16
	4	16	20	20	24	20
	5	20	24	24	28	24

Uniform Distribution $[\mu_{ij}+/- 1]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{2}]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{3}]$	Uniform Distribution $[\mu_{ij}+/- 2]$
Significant Difference (+)	No Significant Difference	No Significant Difference	Significant Difference

Table 6.9: Robustness Check to Misspecification of Data Distribution: Comparison of Mean Average Squared Loss of the Normal vs. Uniform Distribution

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	12	16	16	20	16
	2	12	12	20	20	16
	3	16	20	20	24	20
	4	16	16	24	24	20
	5	16	20	24	28	22

Uniform Distribution $[\mu_{ij}+/- 1]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{2}]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{3}]$	Uniform Distribution $[\mu_{ij}+/- 2]$
Significant Difference (+)	No Significant Difference	No Significant Difference	Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	20	20	20	20	20
	2	16	20	20	24	20
	3	16	20	24	28	22
	4	20	24	28	32	26
	5	24	28	32	36	30

Uniform Distribution $[\mu_{ij}+/- 1]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{2}]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{3}]$	Uniform Distribution $[\mu_{ij}+/- 2]$
Significant Difference (+)	No Significant Difference	No Significant Difference	Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	12	16	20	24	18
	2	16	20	24	28	22
	3	20	24	28	32	26
	4	24	28	32	36	30
	5	28	32	36	40	34

Uniform Distribution $[\mu_{ij}+/- 1]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{2}]$	Uniform Distribution $[\mu_{ij}+/-\sqrt{3}]$	Uniform Distribution $[\mu_{ij}+/- 2]$
Significant Difference (+)	No Significant Difference	No Significant Difference	Significant Difference

Table 6.10: Robustness Check to Misspecification of Data Distribution: Comparison of Mean Average Squared Loss of the Normal vs. t Distribution

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	16	20	20	24	20
	2	20	20	20	20	20
	3	16	16	24	24	20
	4	12	16	24	28	20
	5	12	20	20	28	20

t Distribution df = 50	t Distribution df = 30	t Distribution df = 20	t Distribution df = 10
No Significant Difference	No Significant Difference	No Significant Difference	No Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	12	16	16	20	16
	2	16	20	20	24	20
	3	20	20	20	20	20
	4	16	20	20	24	20
	5	16	16	24	24	20

t Distribution df = 50	t Distribution df = 30	t Distribution df = 20	t Distribution df = 10
No Significant Difference	No Significant Difference	No Significant Difference	No Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	12	12	20	20	16
	2	12	16	16	20	16
	3	20	20	20	20	20
	4	16	20	20	24	20
	5	16	16	24	24	20

t Distribution df = 50	t Distribution df = 30	t Distribution df = 20	t Distribution df = 10
No Significant Difference	No Significant Difference	No Significant Difference	Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	12	16	16	20	16
	2	16	16	16	16	16
	3	12	12	20	20	16
	4	16	20	20	24	20
	5	20	24	24	28	24

t Distribution df = 50	t Distribution df = 30	t Distribution df = 20	t Distribution df = 10
No Significant Difference	No Significant Difference	No Significant Difference	Significant Difference

Table 6.11: Robustness Check to Misspecification of Data Distribution: Comparison of Mean Average Squared Loss of the Normal vs. t Distribution

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	12	16	16	20	16
	2	12	12	20	20	16
	3	16	20	20	24	20
	4	16	16	24	24	20
	5	16	20	24	28	22

t Distribution df = 50	t Distribution df = 30	t Distribution df = 20	t Distribution df = 10
No Significant Difference	No Significant Difference	No Significant Difference	Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	20	20	20	20	20
	2	16	20	20	24	20
	3	16	20	24	28	22
	4	20	24	28	32	26
	5	24	28	32	36	30

t Distribution df = 50	t Distribution df = 30	t Distribution df = 20	t Distribution df = 10
No Significant Difference	No Significant Difference	No Significant Difference	No Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	12	16	20	24	18
	2	16	20	24	28	22
	3	20	24	28	32	26
	4	24	28	32	36	30
	5	28	32	36	40	34

t Distribution df = 50	t Distribution df = 30	t Distribution df = 20	t Distribution df = 10
No Significant Difference	No Significant Difference	No Significant Difference	Significant Difference

Table 6.12: Robustness to Constant Variance Assumption: Comparison of Mean Average Squared Loss of Constant and Pseudo-Constant Observation Variance

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	8	10	10	12	10
	2	10	10	10	10	10
	3	8	8	12	12	10
	4	6	8	12	14	10
	5	6	10	10	14	10

Psuedo-Constant Variance = 1 [0.5, 1.5]	Psuedo-Constant Variance = 2 [1, 3]	Psuedo-Constant Variance = 3 [1, 5]
No Significant Difference	No Significant Difference	No Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	6	8	8	10	8
	2	8	10	10	12	10
	3	10	10	10	10	10
	4	8	10	10	12	10
	5	8	8	12	12	10

Psuedo-Constant Variance = 1 [0.5, 1.5]	Psuedo-Constant Variance = 2 [1, 3]	Psuedo-Constant Variance = 3 [1, 5]
No Significant Difference	No Significant Difference	No Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	6	6	10	10	8
	2	6	8	8	10	8
	3	10	10	10	10	10
	4	8	10	10	12	10
	5	8	8	12	12	10

Psuedo-Constant Variance = 1 [0.5, 1.5]	Psuedo-Constant Variance = 2 [1, 3]	Psuedo-Constant Variance = 3 [1, 5]
No Significant Difference	No Significant Difference	No Significant Difference

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	6	8	8	10	8
	2	8	8	8	8	8
	3	6	6	10	10	8
	4	8	10	10	12	10
	5	10	12	12	14	12

Psuedo-Constant Variance = 1 [0.5, 1.5]	Psuedo-Constant Variance = 2 [1, 3]	Psuedo-Constant Variance = 3 [1, 5]
No Significant Difference	No Significant Difference	No Significant Difference

Table 6.13: Robustness to Constant Variance Assumption: Comparison of Mean Average Squared Loss of Constant and Pseudo-Constant Observation Variance

	Individual (j)				
Group (i)	1	2	3	4	Group Mean
1	6	8	8	10	8
2	6	6	10	10	8
3	8	10	10	12	10
4	8	8	12	12	10
5	8	10	12	14	11

Psuedo-Constant Variance = 1 [0.5, 1.5]	Psuedo-Constant Variance = 2 [1, 3]	Psuedo-Constant Variance = 3 [1, 5]
No Significant Difference	No Significant Difference	No Significant Difference

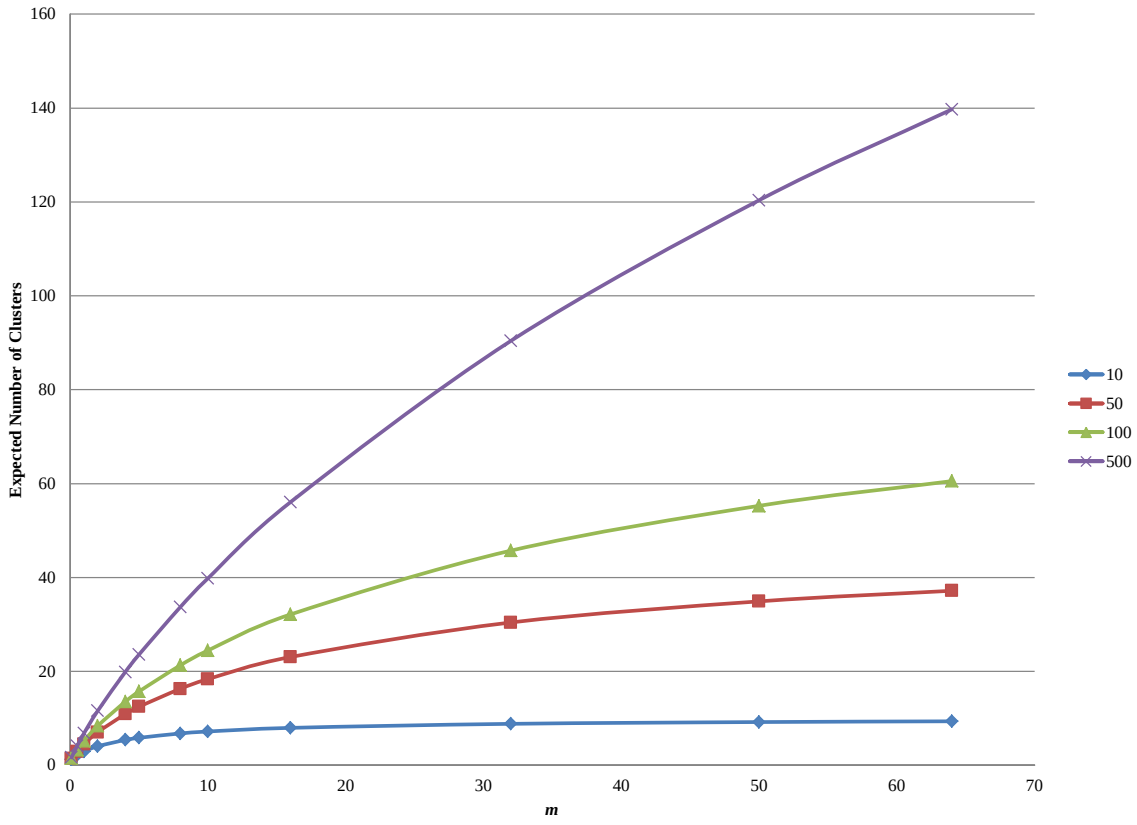
	Individual (j)				
Group (i)	1	2	3	4	Group Mean
1	10	10	10	10	10
2	8	10	10	12	10
3	8	10	12	14	11
4	10	12	14	16	13
5	12	14	16	18	15

Psuedo-Constant Variance = 1 [0.5, 1.5]	Psuedo-Constant Variance = 2 [1, 3]	Psuedo-Constant Variance = 3 [1, 5]
No Significant Difference	No Significant Difference	No Significant Difference

	Individual (j)				
Group (i)	1	2	3	4	Group Mean
1	6	8	10	12	9
2	8	10	12	14	11
3	10	12	14	16	13
4	12	14	16	18	15
5	14	16	18	20	17

Psuedo-Constant Variance = 1 [0.5, 1.5]	Psuedo-Constant Variance = 2 [1, 3]	Psuedo-Constant Variance = 3 [1, 5]
No Significant Difference	No Significant Difference	No Significant Difference

Figure 6.3: Expected Number of Clusters by Total Number of Items and Values of m



7 Comparison of Prior Partition Distributions

Thus far, the prior partition distribution implemented by [Cro97], also known more generally as the Ewens-Pitman prior [CMG$^+$14], has been used for both the group and individual level partitions. In the PPM literature, this and similar priors are the most popular, but are hardly the only ones used. As noted in [QI03], the choice of prior distribution for the partition has a substantial impact on the results. Discussed herein are the Uniform Prior, the Constant Prior, the Hierarchical Uniform Prior, and Ewens-Pitman Prior. In the two-level setting, partitions at the group and individual-levels must be considered. As such, there is the opportunity for setting multiple priors, not necessarily of the same type.

7.1 Uniform and Constant Priors

One of the easiest to concieve, and thus frequently used, prior partition distributions is the uniform prior. With this distribution each of the possible partitions is given the same prior probability. As noted before, for n items the number of partitions with s sets or clusters is a Stirling number of the second kind, $S(n, s)$. Considering

that s may equal $1, 2, ..., n$, the total number of possible partitions items is provided by the Bell number, $B_n = \sum_{s=1}^{n} S(n, s)$. As such, giving equal weight to each of the possible partitions of n items, the uniform prior is thus defined as $P^U = \frac{1}{B_n}$. An alternative specification is sometimes referred to as the constant prior, $P^C \propto m^s$ where m is a hyperparameter which must be estimated or assigned. [QI03, JLB07]

7.2 Hierarchical Uniform Prior

Casella, Morena, and Giron (2014) [CMG+14] introduces the "hierarchical uniform prior" through a decomposition. The component pieces together consider and condition upon what the authors term the *configuration classes* of the partition space given the number of clusters s. These configuration classes are determined by how the items are assigned to the s sets constituting the partition.

Letting n_s be the number of items in set s, a particular configuration class is noted by $\mathfrak{R}_{s;n_1,n_2,...,n_s}$ with $n_1 \leq n_2 \leq ... \leq n_s$. The number of configuration classes depends on both n and s. For example, with $n = 5$, the possible partition space includes partitions with $s = 1, ..., 5$ clusters or sets. When s equals 2 or 3, there are two configuration classes, otherwise when there is one configuration class (Table 7.1). As another example, if $n = 3$, there is one configuration class no matter if $s = 1, 2,$ or 3. The number of configuration classes, $b(n, s)$, for n items and s clusters may be calculated through the recursive expression:

$$b(n, s) = b(n - 1, s - 1) + b(n - s, s)$$

where $b(n, s) = 0$ if $s > n$, and $b(n, s = 1)$ and $b(n, s = n) = 1$. Alternatively,

as $n \to \infty$, $b(n,s) \approx \frac{n^{s-1}}{s!(s-1)!}$ ([WGC14]). This number can then be used to calculate the number of possible partitions with s sets, equivalently $S(n,s)$. It can be shown that the number of partitions in configuration class $\mathfrak{R}_{s;n_1,n_2,...,n_s} = \binom{n}{n_1...n_s}\frac{1}{R(n_1,...,n_s)}$ where $R(n_1,...,n_s) = \prod_{i=1}^{n}\Big(\sum_{k=1}^{s} I(n_k = i)\Big)!$. Thus, $S(n,s) = \sum\binom{n}{n_1...n_s}\frac{1}{R(n_1,...,n_s)}$ where the summation is over all arrangments of the n items into s sets under the condition that $n_1 \leq n_2 \leq ... \leq n_s$.

For a partition ρ with s sets, decomposing the partition distribution to condition upon these configuration classes yields:

$$P^{HUP} = P(\rho|\mathfrak{R}_{s;n_1,n_2,...,n_s}, s)P(\mathfrak{R}_{s;n_1,n_2,...,n_s}|s)P(s)$$

The authors make the assumption that assuming s sets, each configuration class in $\mathfrak{R}_s$, the set of partitions with s clusters such that

$$\mathfrak{R}_s = \bigcup_{\substack{n_1 \leq n_2 \leq ... \leq n_s \\ \sum_{i=1}^{s} n_i = n}} \mathfrak{R}_{s;n_1,n_2,...,n_s},$$

is equally likely. As such,

$$P(\mathfrak{R}_{s;n_1,n_2,...,n_s}|s) = b(n,s)^{-1}$$

Continuing, conditioning on configuration class $P(\mathfrak{R}_{n_1,n_2,...,n_s}|s)$ and s, each partition in $\mathfrak{R}_{n_1,n_2,...,n_s}$ is also equally likely. Thus,

$$P(\rho|\mathfrak{R}_{s;n_1,n_2,...,n_s}) = \binom{n}{n_1...n_s}^{-1} R(n_1,...,n_s)$$

So, $P^{HUP} = \binom{n}{n_1 \dots n_s}^{-1} R(n_1, ..., n_s) b(n, s)^{-1} P(s)$. The authors specify $P(s)$ by utilizing a truncated ($s = 1, ..., n$) Poisson distribution, $\text{Poiss}_n(s|\lambda)$, a hyperprior on λ, and integrating away λ. Two hyperpriors are discussed: (1) the improper Jeffreys prior $P(\lambda) \propto \lambda^{-1/2}$, and (2) a prior $P(\lambda)$ provided by [Mor05] where the latter is chosen due to the fact that it provides desired *a priori* smaller probabilities as s approaches n than does the Jeffreys. In otherwords, a more severe penalty is imposed for having many small clusters as $s \to n$. As n grows large, this may be a more and more attractive penalty. As such,

$$P(s) = \frac{\int_0^\infty \frac{\lambda^s e^{-\lambda}}{s!} P(\lambda|\lambda_0{=}1) d(\lambda)}{\sum_{s=1}^n \int_0^\infty \frac{\lambda^s e^{-\lambda}}{s!} P(\lambda|\lambda_0{=}1) d(\lambda)}, \qquad s = 1, ..., n$$

Thus, the complete partition prior is

$$P^{HUP} = \frac{R(n_1,...,n_s)}{b(n,s)\binom{n}{n_1 \dots n_s}} \frac{\int_0^\infty \frac{\lambda^s e^{-\lambda}}{s!} P(\lambda|\lambda_0{=}1) d(\lambda)}{\sum_{s=1}^n \int_0^\infty \frac{\lambda^s e^{-\lambda}}{s!} P(\lambda|\lambda_0{=}1) d(\lambda)},$$

A considerably simpler HUP partition prior arises if there is no need for penalty as $s \to n$ as in [MGVP16] by setting a uniform prior on s, $P(s) = 1/n$, yielding:

$$P^{HUP}_{\text{no penalty}} = \frac{R(n_1,...,n_s)}{n b(n,s)\binom{n}{n_1 \dots n_s}}$$

Though there is a desire to avoid *encouraging* many singleton clusters (ex. setting or preferencing a large m_{gr} value), over-penalizing separation when items should be separate may be particularly harmful at the group-level of the two-level setting. As such, the latter prior, $P^{HUP}_{\text{no penalty}}$, is herein considered.

7.3 Ewens-Pitman Prior

In varying forms and frequently not by this name, the most popular prior in the PPM literature is the Ewens-Pitman prior. This is due to its connection to nonparametric Bayesian models with Dirichlet process priors as shown by [QI03] and outlined below as well as in numerous sources (see for example [D$^+$09] and [MQ10]). The derivation proceeds by assuming the following model of i.i.d. draws following F where F is a Dirichlet Process:

$$\xi_1, \xi_2, ..., \xi_n \overset{i.i.d.}{\sim} F$$

$$F \sim DP(m, F_0)$$

where m is a weight parameter and F_0 is the centering probability measure of the Dirichlet Process with $E(F) = F_0$. It can be shown that through integration over the Dirichlet Process the resulting distribution over the $\xi_1, \xi_2, ..., \xi_n$ draws follows a Pólya urn scheme:

$$P(\xi_1, \xi_2, ..., \xi_n) = \prod_{i=1}^{n} \left\{ \frac{mF_0(\xi_i) + \sum_{j<i} \delta_{\xi_j}(\xi_i)}{m + i - 1} \right\}$$

such that $\delta_{\xi_j}(\xi_i) = 1$ if $\xi_j = \xi_i$ and 0 otherwise. By the nature of the Dirichlet Process, there is a positive probability of draws sharing the same value. Accordingly, define a partition $\rho = (S_1, ..., S_s)$ as earlier where each S_j is a non-overlapping subset of the n draws such that all of the draws in S_j share the same value. [QI03] makes use of the results in [L$^+$84] to provide an alternative expression for $P(\xi_1, \xi_2, ..., \xi_n)$ which implies that

$$P^{EP}(\rho, s) = \frac{\Gamma(m)}{\Gamma(n + m)} \prod_{k=1}^{s} m(n_{S_k} - 1)!$$

where n_{S_k} is the number of elements in S_k.

As discussed, m must be estimated, assigned, or otherwise considered. In Chapter 3, it is shown that this prior is equivalent to that used by [Cro97] and is used explicitly in many of the cited sources expressed above and throughout this dissertation.

Table 7.1: Decomposition of $n = 5$ items into possible partitions with $s = 1, 2, ..., 5$ clusters by configuration class

$s = 1$	$s = 2$		$s = 3$		$s = 4$	$s = 5$
Configuration Classes						
$\mathfrak{R}_{1;5}$ $n_1 = 5$	$\mathfrak{R}_{2;1,4}$ $n_1 = 1$ $n_2 = 4$	$\mathfrak{R}_{2;2,3}$ $n_1 = 2$ $n_2 = 3$	$\mathfrak{R}_{3;1,1,3}$ $n_1 = 1$ $n_2 = 1$ $n_3 = 3$	$\mathfrak{R}_{3;1,2,2}$ $n_1 = 1$ $n_2 = 2$ $n_3 = 2$	$\mathfrak{R}_{4;1,1,1,2}$ $n_1 = 1$ $n_2 = 1$ $n_3 = 1$ $n_4 = 2$	$\mathfrak{R}_{5;1,1,1,1,1}$ $n_1 = 1$ $n_2 = 1$ $n_3 = 1$ $n_4 = 1$ $n_5 = 1$
(1,2,3,4,5)	(1)(2,3,4,5) (2)(1,3,4,5) (3)(1,2,4,5) (4)(1,2,3,5) (5)(1,2,3,4)	(4,5)(1,2,3) (3,5)(1,2,4) (3,4)(1,2,5) (2,5)(1,3,4) (2,4)(1,3,5) (2,3)(1,4,5) (1,5)(2,3,4) (1,4)(2,3,5) (1,3)(2,4,5) (1,2)(3,4,5)	(4)(5)(1,2,3) (3)(5)(1,2,4) (3)(4)(1,2,5) (2)(5)(1,3,4) (2)(4)(1,3,5) (2)(3)(1,4,5) (1)(5)(2,3,4) (1)(4)(2,3,5) (1)(3)(2,4,5) (1)(2)(3,4,5)	(5)(1,2)(3,4) (4)(1,2)(3,5) (3)(1,2)(4,5) (5)(1,3)(2,4) (4)(1,3)(2,5) (2)(1,3)(4,5) (5)(1,4)(2,3) (3)(1,4)(2,5) (2)(1,4)(3,5) (2)(1,5)(3,4) (4)(1,5)(2,3) (3)(1,5)(2,4) (1)(2,3)(4,5) (1)(2,4)(3,5) (1)(2,5)(3,4)	(3)(4)(5)(1,2) (2)(4)(5)(1,3) (2)(3)(5)(1,4) (2)(3)(4)(1,5) (1)(4)(5)(2,3) (1)(3)(5)(2,4) (1)(3)(4)(2,5) (1)(2)(5)(3,4) (1)(3)(4)(3,5) (1)(2)(3)(4,5)	(1)(2)(3)(4)(5)

Simulation

The comparisons herein focus on results under a changing group-level prior partition distribution. As noted earlier, in the one-level PPM setting, priors equivalent or similar to the Ewens-Pitman partition prior are generally the most popular in the literature. In keeping with this precedent, the individual-level partition prior is specified as a Ewens-Pitman prior with empirically set m_{ind}. As this prior is so often used, it is natural to compare the other prior distributions against this standard at the group-level. Tables 7.2 - 7.3 provide the difference in mean average squared loss of results stemming from specifications with uniform, constant, and hierarchical uniform group-level partition priors with those from a specification with a Ewens-Pitman prior with m_{gr} set to 1 as in the previous chapter. There are several notable results. The first is that the comparisons with the uniform and constant priors yield quite similar results, as expected. Secondly, though the difference may be small, the a Ewens-Pitman group-level prior generally yields superior results to the uniform, constant, and HUP priors as evidenced by the negative differences for the settings reviewed. Thirdly, this advantage seems to dissipate as the number of unequal group means increases. The benefit of the Ewens-Pitman prior is not necessarily lost simply by adding more groups however. For example in Table 7.4, increasing g from 5 to 7, using a Ewens-Pitman prior at the group-level still results in superior estimates to those of the others.

Table 7.2: Difference in Mean Average Squared Loss by Prior Partition Distributions, 30 simulations/comparison, $\sigma^2 = 1$

	Individual (j)				
Group (i)	1	2	3	4	Group Mean
1	16	20	20	24	20
2	20	20	20	20	20
3	16	16	24	24	20
4	12	16	24	28	20
5	12	20	20	28	20

Ewens-Pitman vs. Uniform Group Partition Prior	Ewens-Pitman vs. Constant Group Partition Prior	Ewens-Pitman vs. HUP Group Partition Prior
-0.056912	-0.055049	-0.004715

	Individual (j)				
Group (i)	1	2	3	4	Group Mean
1	12	16	16	20	16
2	16	20	20	24	20
3	20	20	20	20	20
4	16	20	20	24	20
5	16	16	24	24	20

Ewens-Pitman vs. Uniform Group Partition Prior	Ewens-Pitman vs. Constant Group Partition Prior	Ewens-Pitman vs. HUP Group Partition Prior
-0.050100	-0.038038	-0.077465

	Individual (j)				
Group (i)	1	2	3	4	Group Mean
1	12	12	20	20	16
2	12	16	16	20	16
3	20	20	20	20	20
4	16	20	20	24	20
5	16	16	24	24	20

Ewens-Pitman vs. Uniform Group Partition Prior	Ewens-Pitman vs. Constant Group Partition Prior	Ewens-Pitman vs. HUP Group Partition Prior
-0.019195	-0.017065	-0.065320

	Individual (j)				
Group (i)	1	2	3	4	Group Mean
1	12	16	16	20	16
2	16	16	16	16	16
3	12	12	20	20	16
4	16	20	20	24	20
5	20	24	24	28	24

Ewens-Pitman vs. Uniform Group Partition Prior	Ewens-Pitman vs. Constant Group Partition Prior	Ewens-Pitman vs. HUP Group Partition Prior
-0.024276	-0.021597	-0.048165

Table 7.3: Difference in Mean Average Squared Loss by Prior Partition Distributions, 30 simulations/comparison, $\sigma^2 = 1$

	Individual (j)				
Group (i)	1	2	3	4	Group Mean
1	12	16	16	20	16
2	12	12	20	20	16
3	16	20	20	24	20
4	16	16	24	24	20
5	16	20	24	28	22

Ewens-Pitman vs. Uniform Group Partition Prior	Ewens-Pitman vs. Constant Group Partition Prior	Ewens-Pitman vs. HUP Group Partition Prior
0.014107	0.014961	0.015771

	Individual (j)				
Group (i)	1	2	3	4	Group Mean
1	20	20	20	20	20
2	16	20	20	24	20
3	16	20	24	28	22
4	20	24	28	32	26
5	24	28	32	36	30

Ewens-Pitman vs. Uniform Group Partition Prior	Ewens-Pitman vs. Constant Group Partition Prior	Ewens-Pitman vs. HUP Group Partition Prior
-0.001145	-0.001794	0.008568

	Individual (j)				
Group (i)	1	2	3	4	Group Mean
1	12	16	20	24	18
2	16	20	24	28	22
3	20	24	28	32	26
4	24	28	32	36	30
5	28	32	36	40	34

Ewens-Pitman vs. Uniform Group Partition Prior	Ewens-Pitman vs. Constant Group Partition Prior	Ewens-Pitman vs. HUP Group Partition Prior
0.000835	0.000947	-0.001255

Table 7.4: Difference in Mean Average Squared Loss by Prior Partition Distributions, 30 simulations/comparison, $\sigma^2 = 1$

		Individual (j)				
		1	2	3	4	Group Mean
Group (i)	1	30	40	40	50	40
	2	40	40	40	40	40
	3	30	30	50	50	40
	4	40	50	50	60	50
	5	50	50	50	50	50
	6	40	50	50	60	50
	7	40	40	60	60	50

Ewens-Pitman vs. Uniform Group Partition Prior	Ewens-Pitman vs. Constant Group Partition Prior	Ewens-Pitman vs. HUP Group Partition Prior
-0.02492436	-0.02411165	-0.02573971

8 Applied Data

Crowley (1997) [Cro97] analyzes batting averages for major league baseball teams. This analysis uses the first 300 at bats of a season (irrespective of player) as the basis for team batting averages. These averages are used as the data observations from which a product estimate is estimated for each team. This figure is used as a prediction of its respective team's batting average for the remainder of the season. The assumption is made that the averages are derived from hits which follow a binomial distribution such that if b_i is the batting average for team i, then $300b_i \sim \text{bin}(300, p_i)$ where p_i is thought of as the batting average for the remainder of the season, or more generally as the underlying true batting average for team i. [Cro97] deals with the clustering problem for data following a Normal distribution with known and constant variance. Accordingly, a variance-stabilizing transformation is implemented such that $X_i = f(b_i) = \sqrt{300}\arcsin(2b_i - 1)$ and $\mu_i = f(p_i) = \sqrt{300}\arcsin(2p_i - 1)$, so $X_i|\mu_i \sim N(\mu_i, 1)$. This transformation is not without precedent. This same transformation is used in [Geo86] on the same data for the same purpose of achieving normality with constant variance 1 for each data observation.

Obtaining the team batting averages data used in [Cro97] (Table 8.1), the computation of a one-level product estimate confirms the squared error loss reported in [Cro97] (reported 4.30, confirmed at 4.31). Each team was part of a larger division,

American League East, American League West, National League East, and National League West as noted in Table 8.1. The division membership may be thought of in the two-level setting as the group membership, and each team may be considered an individual-level item. As in [Cro97], squared error loss may be computed comparing the batting average estimates with the known batting averages for the remainder of the season.

Table 8.1: 1984 Major League Baseball Team Batting Averages

ID	Group Level Index	Individual Level Index	Division	Team	Team Batting Average after 300 At Bats	Total Season Team Batting Average
1	1	1	American League East	Boston	0.239	0.284
2	1	2	American League East	New York	0.244	0.277
3	1	3	American League East	Toronto	0.283	0.273
4	1	4	American League East	Detroit	0.278	0.271
5	1	5	American League East	Cleveland	0.256	0.265
6	1	6	American League East	Milwaukee	0.272	0.262
7	1	7	American League East	Baltimore	0.247	0.253
8	2	1	American League West	Kansas City	0.279	0.268
9	2	2	American League West	Minnesota	0.280	0.264
10	2	3	American League West	Texas	0.238	0.262
11	2	4	American League West	Oakland	0.244	0.260
12	2	5	American League West	Seattle	0.271	0.256
13	2	6	American League West	California	0.228	0.250
14	2	7	American League West	Chicago	0.222	0.249
15	3	1	National League East	Philadelphia	0.287	0.265
16	3	2	National League East	Chicago	0.284	0.259
17	3	3	National League East	New York	0.248	0.258
18	3	4	National League East	Pittsburgh	0.238	0.255
19	3	5	National League East	St. Louis	0.285	0.250
20	3	6	National League East	Montreal	0.294	0.248
21	4	1	National League West	San Francisco	0.260	0.265
22	4	2	National League West	Houston	0.229	0.265
23	4	3	National League West	San Diego	0.284	0.258
24	4	4	National League West	Atlanta	0.194	0.250
25	4	5	National League West	Los Angeles	0.244	0.244
26	4	6	National League West	Cincinnati	0.265	0.242

Running the two-level analysis as described earlier in this dissertation, $m_{\text{gr}} = 1$ and m_{ind} set empirically (interval width equal to 1), yields improved results with the two-level analysis achieving a squared error loss of 4.18, less than not just the one-

Table 8.2: Squared Error Loss for Season Batting Average by Estimate Type

Estimate Type	Squared Error Loss
Two-Level Product Estimate	**4.18**
One-Level Product Estimate	4.30
MLE	26.64
Empirical Bayes	4.28
Mixture Model	4.33
Best of George (1986)'s general Stein Estimates	4.24
George (1986) Multiple Shrinkage Estimates	5.37, 4.73, and 4.43

level product estimate, but also superior to the other estimates reported in [Cro97] as shown in Table 8.2.

For future analysis, an interesting further application comes if player-level data is gathered. This may be done by identifying the date in the season where the data in Table 8.1 was obtained and obtaining each player's batting average from their most recent 300 at bats (one would have to gather data from the previous season as well). The player-level averages could then be treated as the individual-level data. Straightforwardly, the teams could be considered as the group-level items. This incorporates a finer detail of data and should provide greater insight into team batting averages.

Incorporating player-level data provides an additional and possibly more interesting application of the two-level setting through tailoring the two-level mean condition to specific situations. For example, regarding batting lineups, the players who bat earlier in the lineup tend to constitute a greater proportion of the team at bats than do players towards the end of the lineup. It is easily possible to go back through the course of several seaons to get average proportions for each of the nine batting positions. The mean condition could be altered such that the group-level means could equal an appropriately weighted average of the individual-level means

rather than a simple equal weights average. This would provide an information rich environment from which to predict a team's season batting average, but also the ability to see how modifying lineup order (where specific players bat in the lineup) may affect predicted overall performance.

9 Directions for Future Research

9.1 Multivariate Setting

The model may be extended to a multivariate setting by restructuring the dataset and associated components as follows. Assume the setting has v variables. Let $\mathbf{Y}_R$ be a restructured dataset, organized such that the data may be presented as a vector with each observation having a triple subscript: ij, v. As before, i provides an index for the group and j provides an index of individuals within a group. The additional index, v, provides a variable index. As such the data may be thought of in this fashion as:

$$\mathbf{Y}_R = \begin{bmatrix} \mathbf{Y}_{11} \\ \mathbf{Y}_{12} \\ \vdots \\ \mathbf{Y}_{1n_1} \\ \mathbf{Y}_{21} \\ \vdots \\ \mathbf{Y}_{gn_g} \end{bmatrix}$$

where each element $\mathbf{Y}_{ij},\ i = 1, ..., g;\ j = 1, ..., n_i$, is a length v vector:

$$\mathbf{Y}_{ij} = \begin{bmatrix} \mathrm{Y}_{ij,1} \\ \mathrm{Y}_{ij,2} \\ \vdots \\ \mathrm{Y}_{ij,\mathrm{v}} \end{bmatrix}$$

Every $\mathbf{Y}_{ij}$ has an associated mean vector $\boldsymbol{\mu}_{ij}$ such that

$$\boldsymbol{\mu}_{ij} = \begin{bmatrix} \mu_{ij,1} \\ \mu_{ij,2} \\ \vdots \\ \mu_{ij,\mathrm{v}} \end{bmatrix}$$

and the vector of individual-level item means is

$$\boldsymbol{\mu}_R = \begin{bmatrix} \boldsymbol{\mu}_{11} \\ \boldsymbol{\mu}_{12} \\ \vdots \\ \boldsymbol{\mu}_{1n_1} \\ \boldsymbol{\mu}_{21} \\ \vdots \\ \boldsymbol{\mu}_{gn_g} \end{bmatrix}$$

Items are mapped into partition sets at group and individual-levels as before. Given the individual-level partition, ρ_{ind}, the group-level partition, ρ_{g}, and the two-level mean condition, a set of free individual-level set means (vectors) may be determined and denoted $\boldsymbol{\mu}_{0R}$. There is an associated mapping matrix denoted $\mathbf{H}_R$ which maps

these free set means (vectors) to each individual item index pair $\mathbf{Y}_{ij}$, and hence a mean $\mu_{ij,v}$ to each index triple, $\mathrm{Y}_{ij,v}$.

Let $w = ||\boldsymbol{\mu}_{0R}||$ be the number of free sets, and hence the number of free set means (vectors) as determined by ρ_{ind} and ρ_{g}. Each set mean (vector) in $\boldsymbol{\mu}_{0R}$, $\boldsymbol{\mu}_{0R_u}$, is a vector of v elements, one for each variable:

$$\boldsymbol{\mu}_{0R_u} = \begin{bmatrix} \mu_{0R_{u1}} \\ \mu_{0R_{u2}} \\ \vdots \\ \mu_{0R_{u\mathrm{v}}} \end{bmatrix}$$

where $u = 1, ..., w$ such that

$$\boldsymbol{\mu}_{0R} = \begin{bmatrix} \boldsymbol{\mu}_{0R_1} \\ \boldsymbol{\mu}_{0R_2} \\ \vdots \\ \boldsymbol{\mu}_{0R_w} \end{bmatrix}$$

$\mathbf{H}_R$ provides a mapping of the free individual-level set means (vectors) such that $\boldsymbol{\mu}_R = \mathbf{H}_R\boldsymbol{\mu}_{0R}$ is a $(\sum_{i=1}^{g} n_i) * \mathrm{v}$ length vector. Finding $\mathbf{H}_R$ amounts to finding $\mathbf{H}$ as in the univariate setting since $\mathbf{H}$ is based on the partitions and not the number of variables, and then augmenting $\mathbf{H}$ to accommodate the additional variables. Instead of w columns, now $w * \mathrm{v}$ are necessary since $||\boldsymbol{\mu}_{0R}|| = w * \mathrm{v}$. Similarly, instead of $(\sum_{i=1}^{g} n_i)$ rows, there are $(\sum_{i=1}^{g} n_i) * \mathrm{v}$ to accommodate for the v variables for each individual-level item. For example, for a given pair of partitions ρ_{g} and ρ_{ind}, if when

$v = 1$,

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

then if $v = 2$,

$$\mathbf{H}_R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

As with the single variable setting, let the data follow a multivariate Normal distribution, but such that:

$$\mathbf{Y}_R \sim MVN(\boldsymbol{\mu}_R = \mathbf{H}_R \boldsymbol{\mu}_{0R}, \boldsymbol{\Sigma}_R | \rho_{\text{g}}, \rho_{\text{ind}})$$

where $\mathbf{\Sigma}_R$ is the covariance matrix for $\mathbf{Y}_R$. Since the $\mathbf{Y}_{ij},\ i = 1, ..., g,\ j = 1, ..., n_i$ are conditionally independent given ρ_{ind} and ρ_{g},

$$\mathbf{\Sigma}_R = \begin{bmatrix} \mathbf{\Sigma}_{11} & 0 & \dots & 0 \\ 0 & \mathbf{\Sigma}_{12} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{\Sigma}_{gn_g} \end{bmatrix}$$

Work has begun on the independent variables setting, but more work is necessary including both this setting and non-independent variables settings.

9.1.1 Independent Variables

In the setting where the variables are independent, each $\mathbf{\Sigma}_{ij}$ is a v $\times$ v diagonal matrix such that:

$$\mathbf{\Sigma}_{ij} = \begin{bmatrix} \sigma^2_{ij,1} & 0 & \dots & 0 \\ 0 & \sigma^2_{ij,2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2_{ij,\text{v}} \end{bmatrix}$$

Like $\boldsymbol{\mu}_{ij}$, each $\mathbf{\Sigma}_{ij}$ corresponds to an item. Extending the univariate setting with known and constant variance to the multivariate setting across all variables specifies that all of the $\sigma^2_{ij,v},\ v = 1, ..., \text{v}$, are equal. This constant can be denoted as σ^2. As such, in this case, $\Sigma_R = \sigma^2 \mathbf{I}_R$.

As in the univariate setting, the marginal data distribution may be decomposed into the conditional data distribution and a prior distribution on the free set means

vector:

$$p(\mathbf{Y}_R|\rho_{\text{g}},\rho_{\text{ind}}) = \int p(\mathbf{Y}_R|\boldsymbol{\mu}_R = \mathbf{H}_R\boldsymbol{\mu}_{0R},\rho_{\text{g}},\rho_{\text{ind}})f(\boldsymbol{\mu}_{0R}|\boldsymbol{\theta}_{0R},\boldsymbol{\gamma}_{0R},\rho_{\text{g}},\rho_{\text{ind}})d\boldsymbol{\mu}_{0R} \quad (9.1)$$

As noted in (9.1) let $\boldsymbol{\mu}_{0R}$ have a prior distribution with mean vector

$$\boldsymbol{\theta}_{0R} = \begin{bmatrix} \boldsymbol{\theta}_{R1} \\ \boldsymbol{\theta}_{R2} \\ \vdots \\ \boldsymbol{\theta}_{Rw} \end{bmatrix}$$

where each $\boldsymbol{\theta}_{Ru}$ is a length v vector such that

$$\boldsymbol{\theta}_{\boldsymbol{Ru}} = \begin{bmatrix} \theta_{Ru_1} \\ \theta_{Ru_2} \\ \vdots \\ \theta_{Ru_{\text{v}}} \end{bmatrix}$$

with $u = 1, ..., w$. Similar to the univariate setting, each $\boldsymbol{\theta}_{\boldsymbol{Ru}}$ may be set equal to some $\boldsymbol{\theta}$ where

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_v \end{bmatrix}$$

Thus,

$$\boldsymbol{\theta}_{0R} = \boldsymbol{\theta}\mathbf{1}_w$$

Since the free set means (vector) are independent, the $\boldsymbol{\gamma}_{0R}$ matrix is block diagonal:

$$\boldsymbol{\gamma}_{0R} = \begin{bmatrix} \boldsymbol{\gamma}_{0R1} & 0 & \dots & 0 \\ 0 & \boldsymbol{\gamma}_{0R2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\gamma}_{0Rw} \end{bmatrix}$$

where each $\boldsymbol{\gamma}_{0R\eta},\ \eta = 1, ..., w$ is itself a v $\times$ v covariance matrix for the v variables. In this formulation, and since the variables are independent, $\boldsymbol{\gamma}_{0R}$ is diagonal in this specification.

Together these settings yield similar joint and posterior distributions for $\rho_{\text{g}}, \rho_{\text{ind}}, \mathbf{Y}_R$ as in the univariate setting.

Simulation

The hyperparameter terms may be set or estimated in a similar fashion as in the univariate setting. The non-m hyperparameters are specified variable by variable. Table 9.1 provides an example of results of this extension to the multivariate setting again with $\sigma^2 = 1$.

There is a concern that as the number of variables grows, that the number of clusters will strongly gravitate towards singletons unless proper controls are otherwise specified. For this example, both m_{gr} and m_{ind} are set equal to 1 to encourage a

Table 9.1: Example of Multivariate Two-Level Analysis, $\sigma^2 = 1$

		Generating Means			Two-Level Product Estimate			Posterior Mode Estimate		
i	j	V1	V2	V3	V1	V2	V3	V1	V2	V3
1	1	30	34.65843	45.06757	29.20928	34.00758	45.35838	29.20947	34.00543	45.35854
1	2	40	47.75165	50.8124	39.56292	47.4374	50.96736	39.56282	47.43919	50.96761
1	3	50	60.84486	56.55724	49.91787	60.87111	56.57743	49.91616	60.87295	56.57668
2	1	40	47.75165	50.8124	39.56292	47.4374	50.96736	39.56282	47.43919	50.96761
2	2	40	47.75165	50.8124	39.56292	47.4374	50.96736	39.56282	47.43919	50.96761
2	3	40	47.75165	50.8124	39.56422	47.44129	50.96846	39.56282	47.43919	50.96761
3	1	40	47.75165	50.8124	39.56249	47.44111	50.96795	39.56282	47.43919	50.96761
3	2	50	60.84486	56.55724	49.91519	60.87169	56.5761	49.91616	60.87295	56.57668
3	3	60	73.93807	62.30207	60.26911	74.30746	62.18563	60.2695	74.30671	62.18575
4	1	30	34.65843	45.06757	29.20857	34.00534	45.35842	29.20947	34.00543	45.35854
4	2	60	73.93807	62.30207	60.26911	74.30746	62.18563	60.2695	74.30671	62.18575
4	3	60	73.93807	62.30207	60.26911	74.30746	62.18563	60.2695	74.30671	62.18575

Average of Individual-Level Item Means by Group as Estimated by Posterior Mode Estimate

		V1	V2	V3
→	Group 1	39.56282	47.43919	50.96761
→	Group 2	39.56282	47.43919	50.96761
→	Group 3	49.91616	60.87295	56.57668
→	Group 4	49.91616	60.87295	56.57668

The True Underlying Partition as Implied by Generating Means Implies, by the Two-Level Mean Condition, that $\mu_{1,v} = \mu_{2,v}$ and $\mu_{3,v} = \mu_{4,v}$

relatively small number of sets at both levels. This appears to have been successful considering the results of both the two-level product estimate and the posterior mode estimate. More experimenation with larger numbers of variables is warranted as to the necessary extent and type of controls.

9.2 Other Specifications

In both the univariate and multivariate settings there are additional specifications that may be considered. One example is modelling explicitly non-constant observation variances rather than estimating a range or ranges of variances and using a mid-point as a pseudo-constant variance. Another example is allowing dependent observations. Considering multivariate data, work has begun on investigating how variable selection might be incorporated into a two-level analysis, specifically regarding model comparison and selection and how model complexity penalties might be applied, but there are many avenues and methods to consider. Current work has focused on incorporating the methods of [AM14] and [RD06]. Additionally, it would be interesting to explore models under all specifications with more than two-levels and multi-level mean conditions, including allowing the conditions to change depending on the specific levels of the data being related.

Acknowledgments

There are not adequate words to describe the debt of gratitude I feel towards my advisor, Professor Jeff Holt. The countless hours of guidance, helpful review of this and other documents, and general counseling he provided are more appreciated than he may ever know. He has been my personal Virgil through all of this, and I am forever grateful.

Thank you to my committee members, Professor Karen Kafadar, Professor Nolan Wages, and Professor Gretchen Martinet. Their assistance and encouragement throughout this process including up through the day of this submission has been greatly helpful. Additionally, a very special thank you is due to Professor Dan Spitzner for his very helpful insight and suggestions leading to this document over a great span of time.

Thank you to Advanced Research Computing Services, notably Jackie Huband, for much appreciated computational assistance and advice.

Finally, many thanks are owed to my family, most notably my wife Christina for her ceaseless support, patience, and love throughout this process. I could not ask for a better partner in life, and I will forever work to equal her example. Her strength has been an inspiration through it all, and I cannot thank her enough.

Truly, thank you to all above.

Bibliography

[AM14] Jeffrey L Andrews and Paul D McNicholas. Variable selection for clustering and classification. *Journal of Classification*, 31(2): 136–153, 2014.

[BH92] Daniel Barry and John A Hartigan. Product partition models for change point problems. *The Annals of Statistics*, pages 260–279, 1992.

[CGB+13] Thiago Costa, Michele Guindani, Federico Bassetti, Fabrizio Leisen, and Edoardo M Airoldi. Generalized species sampling priors with latent beta reinforcements. 2013.

[CMG+14] George Casella, Elías Moreno, F Javier Girón, et al. Cluster analysis, model selection, and prior distributions on models. *Bayesian Analysis*, 9(3): 613–658, 2014.

[Cro92] E. M. Crowley. *Estimation of Cluster Parameters.* PhD thesis, Yale University, Department of Statistics, 1992.

[Cro97] Evelyn M Crowley. Product partition models for normal means. *Journal of the American Statistical Association*, 92(437): 192–198, 1997.

[D+09] David B Dahl et al. Modal clustering in a class of product partition models. *Bayesian Analysis*, 4(2): 243–264, 2009.

[DAHH02] DGT Denison, NM Adams, CC Holmes, and DJ Hand. Bayesian par-

tition modelling. *Computational statistics & data analysis*, 38(4): 475–485, 2002.

[DDT16] David B Dahl, Ryan Day, and Jerry W Tsai. Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association*, (just-accepted), 2016.

[DH01] DGT Denison and CC Holmes. Bayesian partitioning for estimating disease risk. *Biometrics*, 57(1): 143–149, 2001.

[FR02] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458): 611–631, 2002.

[FRS12] Chris Fraley, AE Raftery, and L Scrucca. Normal mixture modeling for model-based clustering, classification, and density estimation. *Department of Statistics, University of Washington*, 23: 2012, 2012.

[Geo86] Edward I George. Combining minimax shrinkage estimators. *Journal of the American Statistical Association*, 81(394): 437–445, 1986.

[Har90] John A Hartigan. Partition models. *Communications in statistics-Theory and methods*, 19(8): 2745–2756, 1990.

[HDM99] CC Holmes, DGT Denison, and BK Mallick. Bayesian partitioning for classification and regression. *Manuscript, Imperial College*, 1999.

[HDRM05] CC Holmes, DG T Denison, S Ray, and BK Mallick. Bayesian prediction via partitioning. *Journal of Computational and Graphical Statistics*, 14(4): 811–830, 2005.

[HPS72] Paul-G Hoel, Sidney-C Port, and Charles-J Stone. Introduction to stochastic processes. 1972.

[JLB07] Claire Jordan, Vicki Livingstone, and Daniel Barry. Statistical modelling using product partition models. *Statistical Modelling*, 7(3): 275–295, 2007.

[L+84] Albert Y Lo et al. On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, 12(1): 351–357, 1984.

[LCAV05] RH Loschi, FRB Cruz, and RB Arellano-Valle. Multiple change point analysis for the regular exponential family using the product partition model. *Journal of Data Science*, 3(3): 305–330, 2005.

[LIAV99] RH Loschi, PL Iglesias, and RB Arellano-Valle. Bayesian detection of change points in the chilean stock market. In *Proceedings of the Annual Meeting of American Statistical Association. Section on Bayesian Statistical Science,(Baltimore, MD, USA, 1999)*, pages 160–165, 1999.

[LMI05] R Loschi, Cristiano R Moura, and Pilar L Iglesias. Bayesian analysis for change points in the volatility of latin american emerging markets. *J. Data Sci*, 3: 101–122, 2005.

[MGVP16] E Moreno, FJ Girón, and FJ Vázquez-Polo. Cost-effectiveness analysis for heterogeneous samples. *European Journal of Operational Research*, 254(1): 127–137, 2016.

[Mor05] Elías Moreno. Objective bayesian methods for one-sided testing. *Test*, 14(1): 181–198, 2005.

[MQ10] Peter Müller and Fernando Quintana. Random partition models with regression on covariates. *Journal of statistical planning and inference*, 140(10): 2801–2808, 2010.

[MQR08] Peter Müller, Fernando Quintana, and Gary Rosner. Bayesian clustering

with regression. Technical report, Tech. rep., MD Anderson Cancer Center, Houston, TX, 2008.

[MQR11] Peter Müller, Fernando Quintana, and Gary L Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1): 260–278, 2011.

[PD10] Ju-Hyun Park and David B Dunson. Bayesian generalized product partition model. *Statistica Sinica*, pages 1203–1226, 2010.

[QI03] Fernando A Quintana and Pilar L Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 557–574, 2003.

[RD06] Adrian E Raftery and Nema Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473): 168–178, 2006.

[WGC14] Andrew Womack, Jeff Gill, and George Casella. Product partitioned dirichlet process prior models for identifying substantive clusters and fitted subclusters in social science data, 2014.