# SDS 383D: Exercises 4 – Hierarchical Models

April 20, 2017

*Professor Scott*

**Spencer Woody**

# Problem 1

## Math Tests

We have a model where $y_{ij}$ is the test score of the $j$th student in school $i$, with indices $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, N_i$, so $N_i$ is the sample size for school $i$ and there are $N = \sum_{i=1}^{I}$ total test scores. Let $\lambda = 1/\sigma^2$ and $\gamma = 1/\tau^2$ be the precision parameters. Further, let $y_i = [y_{i1}, y_{i2}, \ldots, y_{iN_i}]^T$ and $y = [y_1^T, y_2^T, \ldots, y_I^T]^T$ and $\theta = [\theta_1, \theta_2, \ldots, \theta_I]^T$. As we can see in Figure 1, schools with smaller sample sizes tend to have more extreme average test scores.

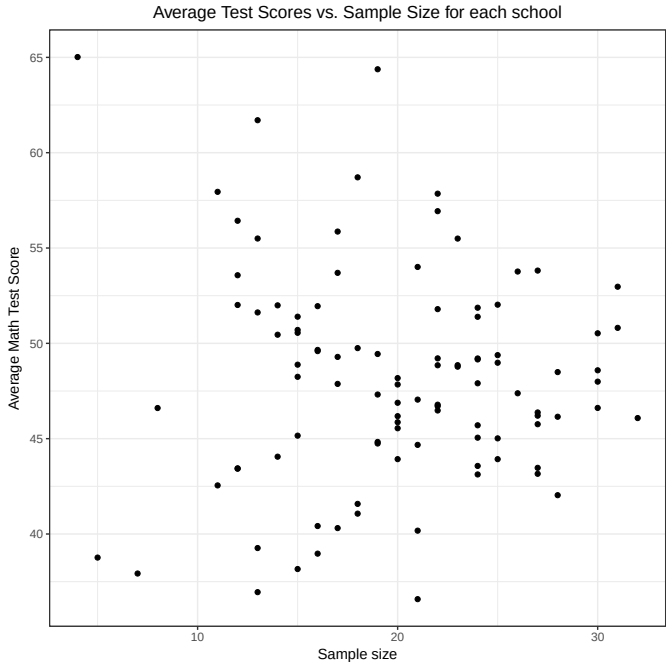


Figure 1: Scatter plot of sample size and average test scores

The hierarchical model for these data is

$$\begin{aligned}(y_{ij}|\theta_i, \lambda) &\sim \mathcal{N}\left(\theta_i, \lambda^{-1}\right) \\ (\theta_i|\mu, \lambda, \gamma) &\sim \mathcal{N}\left(\mu, (\lambda\gamma)^{-1}\right).\end{aligned}$$

We set the priors

$$\begin{aligned}\pi(\mu) &\propto 1, \ -\infty < \mu < \infty \\ \pi(\lambda) &\propto \lambda^{-1}, \ \lambda > 0 \\ \pi(\gamma) &\propto 1, \ \gamma > 0,\end{aligned}$$

that is to say, .... In order to implement the Gibbs sampler, we need the posterior full conditionals for each $\theta_i$, $\mu$, $\lambda$, and $\gamma$.

- For each $\theta_i$,

$$\begin{aligned}f(\theta_i|y_i, \mu, \lambda, \gamma) &\propto f(y_i|\theta_i, \lambda) \cdot f(\theta_i|\mu, \lambda, \gamma) \\ &\sim \mathcal{N}\left((N_i\lambda + \lambda\gamma)^{-1} \cdot (N_i\lambda\bar{y}_i + \lambda\gamma\mu), (N_i\lambda + \lambda\gamma)^{-1}\right),\end{aligned}$$

  which we know from the normal-normal conjugacy derived in Exercises 1.

- For $\mu$,

$$\begin{aligned}
\pi(\mu|y,\theta,\lambda,\gamma) &\propto f(\theta|\lambda,\gamma,\mu)\cdot\pi(\mu) \\
&\propto \left(\prod_{i=1}^{I}\exp\left[-\frac{1}{2}\lambda\gamma(\theta_i-\mu)^2\right]\right)\cdot 1 \\
&= \exp\left[-\frac{1}{2}\lambda\gamma\sum_{i=1}^{I}(\theta_i-\mu)^2\right] \\
&= \exp\left[-\frac{1}{2}\lambda\gamma\sum_{i=1}^{I}\left(\theta_i^2-2\theta_i\mu+\mu^2\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\lambda\gamma\left(I\mu^2-2I\bar{\theta}\mu\right)\right] \\
&\sim \mathcal{N}\left(\bar{\theta},(I\lambda\gamma)^{-1}\right).
\end{aligned}$$

- For $\lambda$,

$$\begin{aligned}
\pi(\lambda|y,\mu,\gamma,\theta) &\propto f(y|\lambda,\theta)\cdot f(\theta|\lambda,\gamma,\mu)\cdot\pi(\lambda) \\
&\propto \left(\prod_{i=1}^{I}\prod_{j=1}^{N_i}\lambda^{1/2}\exp\left[-\frac{1}{2}(y_{ij}-\theta_i)^2\right]\right)\cdot\left(\prod_{i=1}^{I}\lambda^{1/2}\exp\left[-\frac{1}{2}\lambda\gamma(\theta_i-\mu)^2\right]\right)\cdot\lambda^{-1} \\
&= \lambda^{(N+I)/2-1}\exp\left[-\frac{1}{2}\left(\sum_{i=1}^{I}\sum_{j=1}^{N_i}(y_{ij}-\theta_i)^2+\gamma\sum_{i=1}^{I}(\theta_i-\mu)^2\right)\lambda\right] \\
&\sim \text{Gamma}\left(\frac{N+I}{2},\frac{1}{2}\left[\sum_{i=1}^{I}\sum_{j=1}^{N_i}(y_{ij}-\theta_i)^2+\gamma\sum_{i=1}^{I}(\theta_i-\mu)^2\right]\right).
\end{aligned}$$

- For $\gamma$,

$$\begin{aligned}
\pi(\gamma|y,\mu,\lambda,\theta) &\propto f(\theta|\lambda,\gamma,\mu)\cdot\pi(\gamma) \\
&\propto \left(\prod_{i+1}^{I}\gamma^{1/2}\exp\left[-\frac{1}{2}\lambda\gamma(\theta_i-\mu)^2\right]\right)\cdot 1 \\
&= \gamma^{I/2}\exp\left[-\frac{1}{2}\lambda\sum_{i=1}^{I}(\theta_i-\mu)^2\cdot\gamma\right] \\
&\sim \text{Gamma}\left(\frac{I}{2}+1,\frac{1}{2}\lambda\sum_{i=1}^{I}(\theta_i-\mu)^2\right).
\end{aligned}$$

Table 1: 95% posterior credible intervals

| | 2.5% | 50% | 97.5% |
|---|---|---|---|
| $\mu$ | 47.03 | 48.10 | 49.18 |
| $\lambda$ | 0.0111 | 0.0118 | 0.0126 |
| $\gamma$ | 2.43 | 3.49 | 5.03 |

Given the posterior mean $\hat{\theta}_i$ as an estimate of $\theta_i$, define the shrinkage coefficient

$$\kappa_i = \frac{\bar{y}_i - \hat{\theta}_i}{\bar{y}_i},$$

Problem 1 continued on next page...

which is a measure incomplete pooling. Figure 2 shows the absolute shrinkage coefficient for each school as a function of sample size. As sample size increases, the shrinkage decreases because we are gaining precision in estimating the school-level mean $\theta_i$.

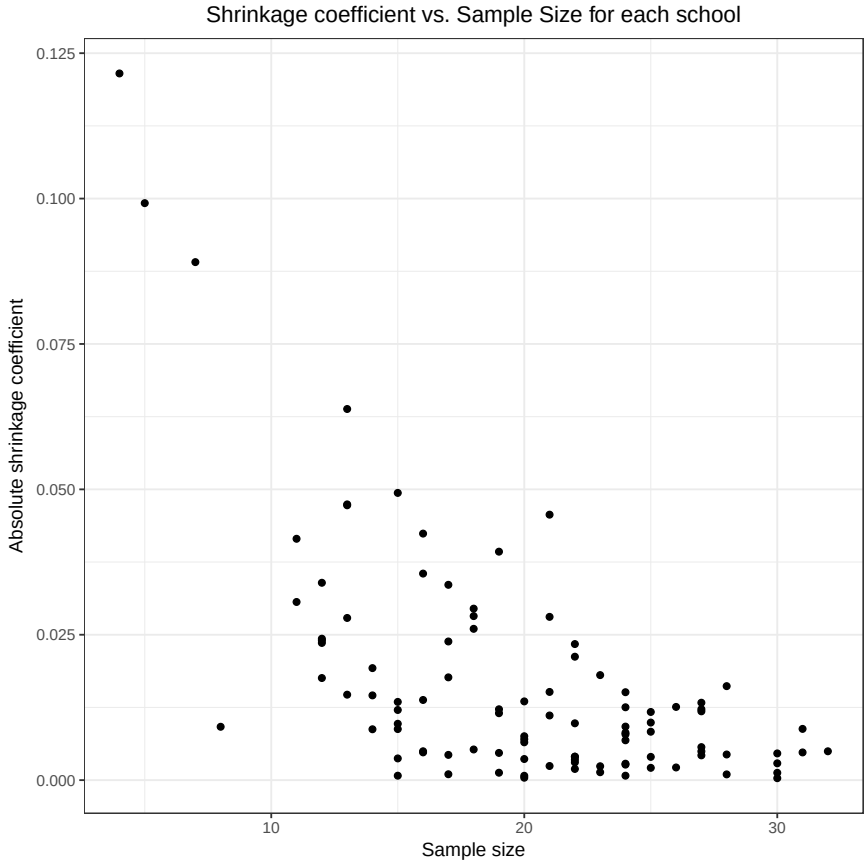


Figure 2: Absolute shrinkage coefficient as a function of sample size

# Problem 2

## Price elasticity of demand

Here we model the demand curve for cheese, which is given by

$$Q = \alpha P^{\beta},$$

where $Q$ is the quantity of cheese demanded, $P$ is price, $\beta$ is a parameter for the *price elasticity of demand* and $\alpha$ is a (rather unremarkable) scaling parameter. Note that if we take a logarithmic transform of the equation in our demand model, we obtain the linear replationship

$$\log Q = \log \alpha + \beta \log P.$$

Figure 3 shows all the data with a fitted OLS line, and Figure 4 shows the data on a store-by-store level with the same OLS line from all data on each panel. The fact that the OLS line performs poorly on any given individual store's data suggests that a hierarchical approach would be beneficial. The hierarchical linear model for the quantity of cheese sold for the $t$th observation at store $i$ is

$$y_{it} = \alpha_i + \beta_i x_{it} + \gamma_i z_{it} + \theta_i z_{it} x_{it} + \epsilon_{it},$$

where $x_{it}$ is the log-price of cheese and $z_{it}$ is an indicator variable taking on a value of 1 when the display is shown, and 0 otherwise.

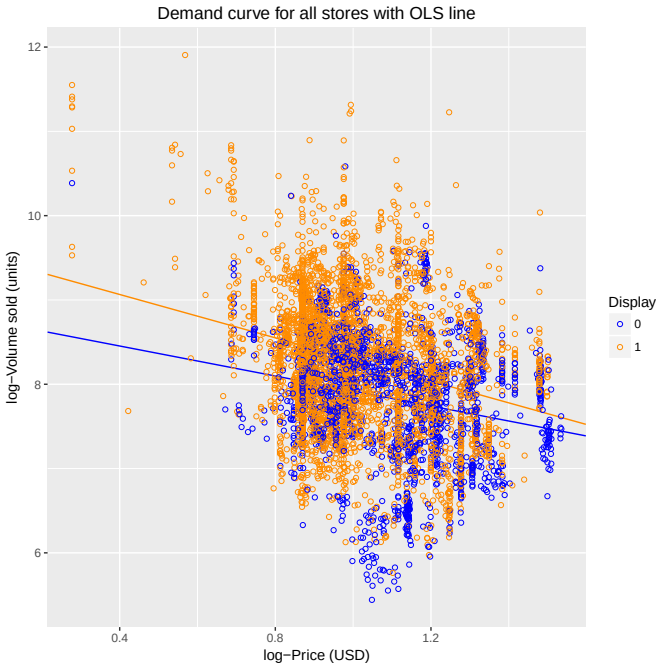


Figure 3: Scatterplot for data from all stores with OLS line

Using freqentist REML to build this model we obtain these results,

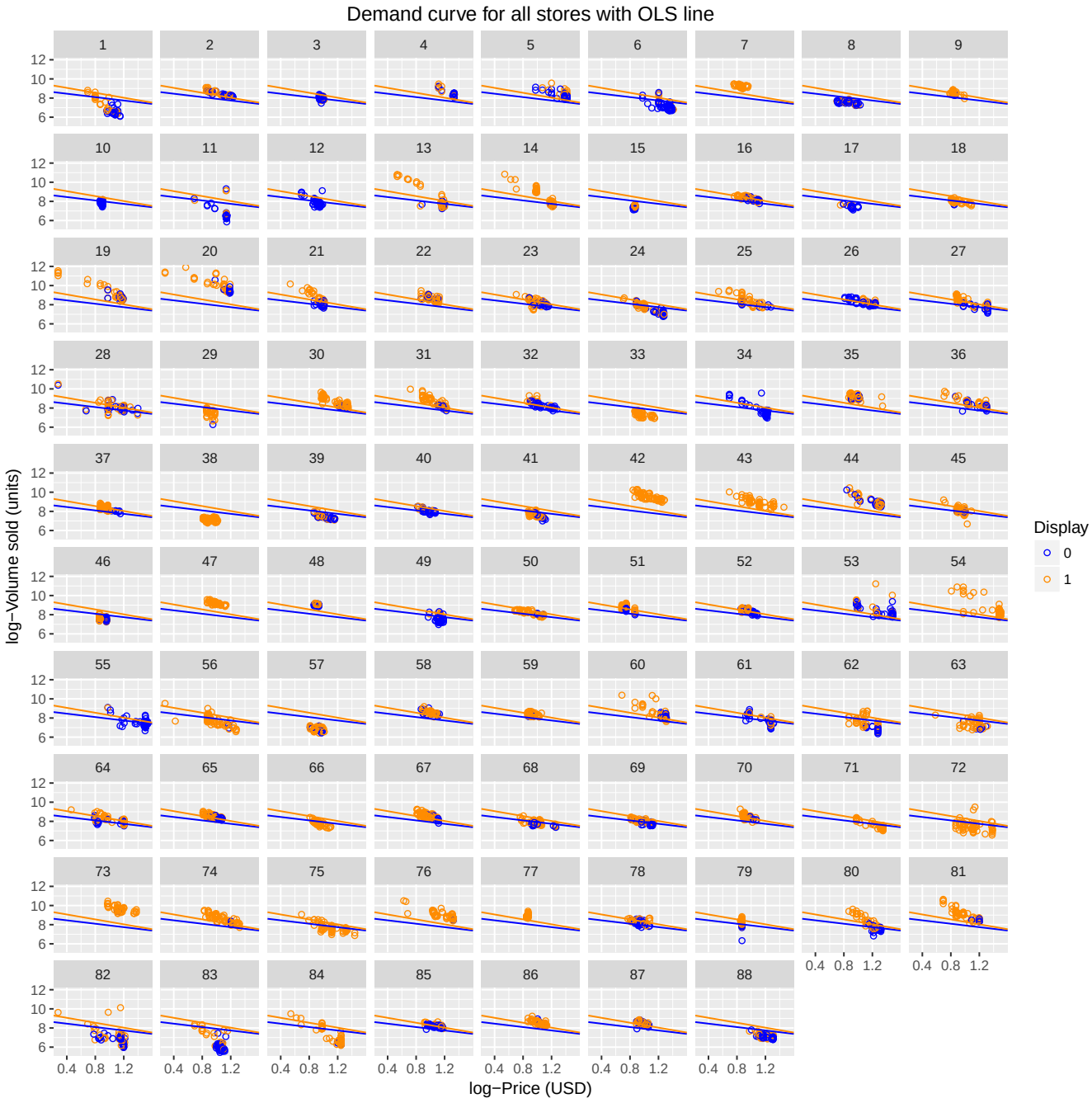


Figure 4: Scatterplot for data from all stores with OLS line

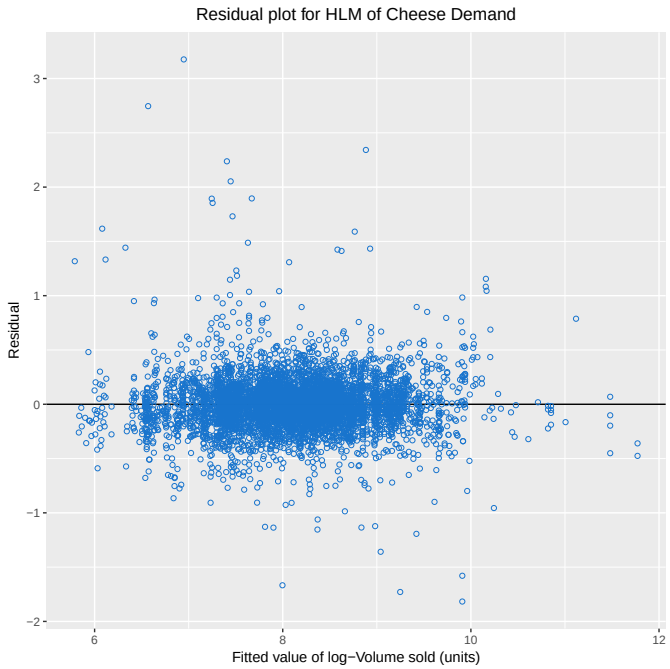


Figure 5: Residual plot using HLM and REML method

*Full Bayesian*

**Model specification**

Here we specify a general Bayesian hierarchical linear model. Let $y_i$ be a $n_i$-length vector representing the the responses of group $i$. There are $N = \sum_i^I n_i$ total responses. $X_i$ is the $n_i \times p$ design matrix for the observations in group $i$, and $Z_i$ is a $n_i \times q$, $q \leq p$ matrix whose columns are a subset of the columns of $X_i$, and this represents the subject-level effects, sometimes called "random effects.". Then the responses $y_i$ are distributed as:

$$\begin{aligned} y_i|\beta, b_i, \lambda &\sim \mathcal{N}_{n_i}(X_i\beta + Z_i b_i, \lambda^{-1}\mathcal{I}_{n_i}) \\ b_i|D &\overset{\text{iid}}{\sim} \mathcal{N}_q(0, D) \end{aligned}$$

Note that the responses $y_{it}$ for subject $i$ are therefore assumed to iid, and also note two results of this model,

$$\begin{aligned} E(y_i|b_i) &= X_i\beta + Z_i b_i \\ E(y_i) &= E(E(y_i|b_i)) = X_i\beta, \end{aligned}$$

or in other words, The priors are

$$\begin{aligned} \pi(\lambda) &\propto \lambda^{-1} \\ \pi(\beta) &\propto 1 \\ \pi(D) &\sim \text{IW}(\nu, \Psi). \end{aligned}$$

To implement a Gibbs sampler, we need the full conditional posterior distributions for $b_i$, $\lambda$, $\beta$, and $D$.

Problem 2 continued on next page... 

- For each $b_i$, first define $v_i := y_i - X_i\beta$,

$$\begin{aligned}
p(b_i|y_i,\lambda,\beta,D) &\propto p(y_i|\beta,b_i,\lambda)p(b_i|D) \\
&\propto \exp\left[-\frac{1}{2}\lambda\left(y_i - X_i\beta - Z_ib_i\right)^T\left(y_i - X_i\beta - Z_ib_i\right)\right]\cdot\exp\left[-\frac{1}{2}b_i^TD^{-1}b_i\right] \\
&= \exp\left[-\frac{1}{2}\lambda\left(Z_ib_i - v_i\right)^T\left(Z_ib_i - v_i\right)\right]\cdot\exp\left[-\frac{1}{2}b_i^TD^{-1}b_i\right] \\
&\propto \exp\left[-\frac{1}{2}b_i^T\left(\lambda Z_i^TZ_i + D^{-1}\right)b_i - 2b_i^T\lambda Z_i^Tv_i\right] \\
&\propto \exp\left[-\frac{1}{2}\left(b_i - \left[\lambda Z_i^TZ_i + D^{-1}\right]^{-1}\lambda Z_i^Tv_i\right)^T\left(\lambda Z_i^TZ_i + D^{-1}\right)\left(b_i - \left[\lambda Z_i^TZ_i + D^{-1}\right]^{-1}\lambda Z_i^Tv_i\right)\right] \\
&\sim \mathcal{N}\left(\left[\lambda Z_i^TZ_i + D^{-1}\right]^{-1}\lambda Z_i^Tv_i, \left[\lambda Z_i^TZ_i + D^{-1}\right]^{-1}\right). \\
&\sim \mathcal{N}\left(\left[\lambda Z_i^TZ_i + D^{-1}\right]^{-1}\lambda Z_i^T(y_i - X_i\beta), \left[\lambda Z_i^TZ_i + D^{-1}\right]^{-1}\right).
\end{aligned}$$

- For $\lambda$,

$$\begin{aligned}
\pi(\lambda|y,\beta,b) &\propto p(y|\lambda,\beta,\underline{\ })\cdot\pi(\lambda) \\
&= \left(\prod_{i=1}^{I}\lambda^{n_i/2}\exp\left[-\frac{1}{2}\lambda(y_i - X_i\beta - Z_ib_i)^T(y_i - X_i\beta - Z_ib_i)\right]\right)\cdot\lambda^{-1} \\
&\sim \text{Gamma}\left(\frac{N}{2}, \frac{1}{2}\sum_{i=1}^{I}\|y_i - X_i\beta - Z_ib_i\|_2^2\right)
\end{aligned}$$

- For $\beta$, define $w_i := y_i - Z_ib_i$.

$$\begin{aligned}
\pi(\beta|y,\lambda,b) &\propto p(y|\lambda,\beta,\underline{\ })\cdot\pi(\beta) \\
&\propto \left(\prod_{i=1}^{I}\exp\left[-\frac{1}{2}\lambda(y_i - X_i\beta - Z_ib_i)^T(y_i - X_i\beta - Z_ib_i)\right]\right)\cdot 1 \\
&= \prod_{i=1}^{I}\exp\left[-\frac{1}{2}\lambda(X_i\beta - w_i)^T(X_i\beta - w_i)\right] \\
&\propto \prod_{i=1}^{I}\exp\left[-\frac{1}{2}\lambda\left(\beta^TX_i^TX_i\beta - 2\beta^TX_i^Tw_i\right)\right] \\
&= \exp\left(-\frac{1}{2}\lambda\left[\beta^T\left(\sum_{i=1}^{I}X_i^TX_i\right)\beta - 2\beta^T\sum_{i=1}^{I}X_i^Tw_i\right]\right) \\
&= \exp\left(-\frac{1}{2}\lambda\left[\beta^T\left(\sum_{i=1}^{I}X_i^TX_i\right)\beta - 2\beta^T\sum_{i=1}^{I}X_i^T(y_i - Z_ib_i)\right]\right) \\
&\sim \mathcal{N}\left(\left[\sum_{i=1}^{I}X_i^TX_i\right]^{-1}\sum_{i=1}^{I}X_i^T(y_i - Z_ib_i), \left[\lambda\sum_{i=1}^{I}X_i^TX_i\right]^{-1}\right).
\end{aligned}$$

Problem 2 continued on next page...

- For $D$,

$$\begin{aligned}\pi(D|b) &\propto p(b|D) \cdot \pi(D) \\ &\propto \left(\prod_{i=1}^{I}[\det(D)]^{-1/2}\exp\left[-\frac{1}{2}b_i^T D^{-1} b_i\right]\right) \cdot [\det(D)]^{-\frac{\nu+q+1}{2}}\exp\left[-\frac{1}{2}\text{tr}(\Psi D^{-1})\right] \\ &\sim \text{IW}\left(I+\nu, \Psi + \sum_{i=1}^{I} b_i b_i^T\right)\end{aligned}$$

The most computationally intensive part of this Gibbs sampler scheme is sampling each $b_i$, and I chose to do this by exploiting a block-diagonal matrix of each $Z_i$ and drawing each $b_i$ simultaneously as a long vector called $b$. For this application specifically, the $X_i$ and $Z_i$ are identical, with a column of 1's for the intercept, a column of log-prices, a column of indicator variables for display, and a column of interaction terms for log-price and display. We run 6000 iterations of the Gibbs sampler with the first 1000 draws discared as burn-in. The `mix` folder within the `img` folder shows traceplots of $\lambda$, each component in $\beta$, and four randomly selected columns of posterior draws of $b$, which all show a good degree of mixing. Histograms for *lambda* and each component of $\beta$ are shown below. Figure 8 shows a grid of plots, each of which has 95% credible intervals of all the subject-level effects on a given covariate terms, arranged in increasing order by posterior median. Note that on the $x$-axis is different for each plot in order to have each one ordered by posterior median.

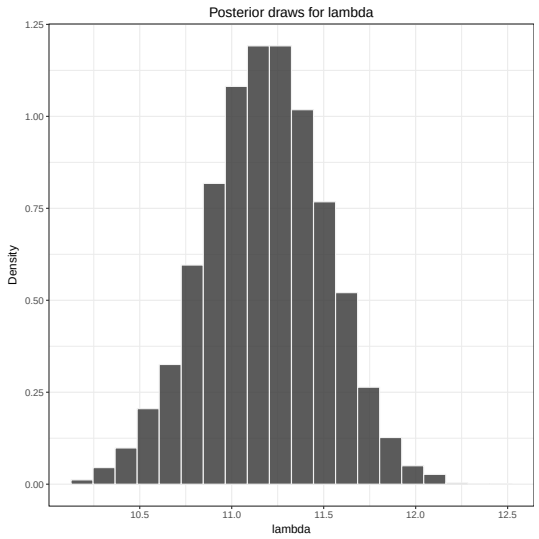


Figure 6: Histogram of posterior draws of $\lambda$

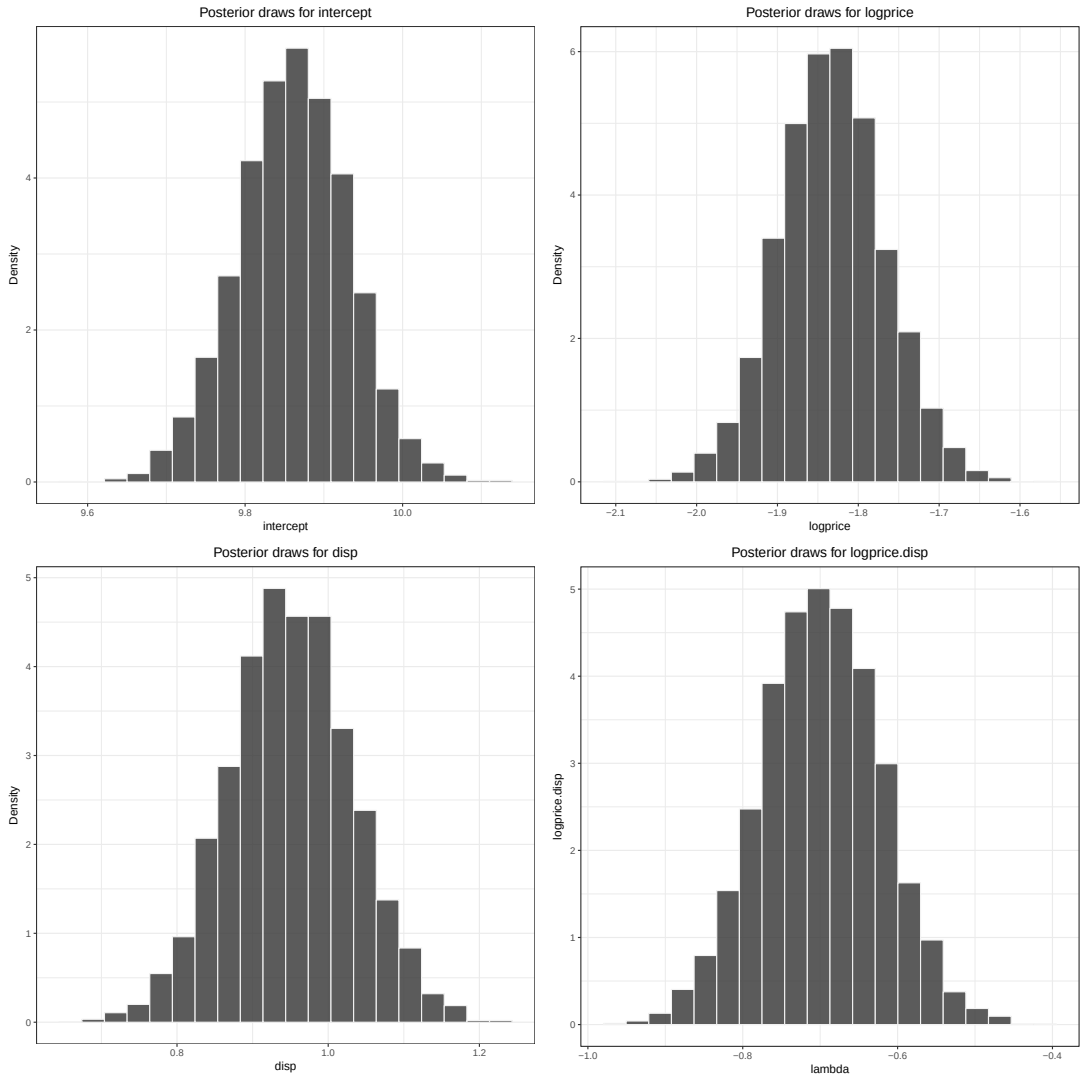


Figure 7: Histogram of posterior draws of each term in $\beta$

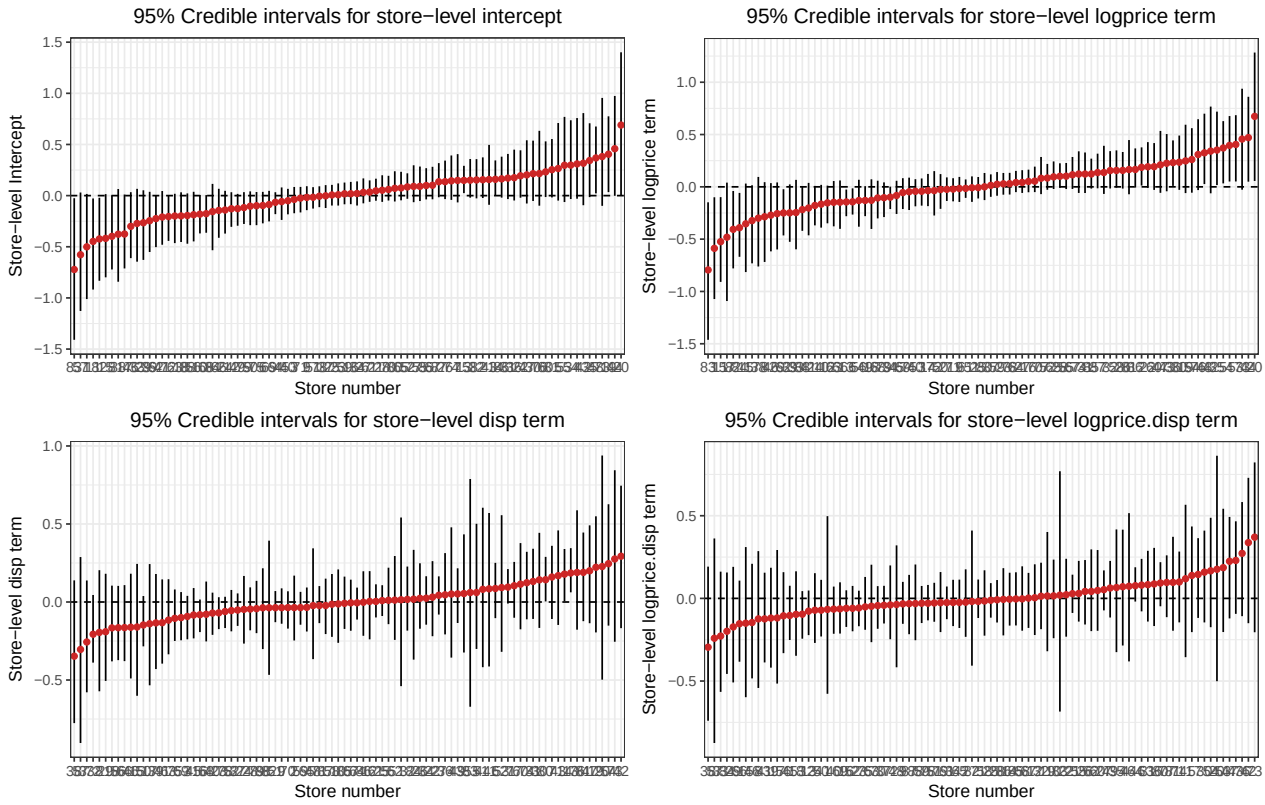


Figure 8: Ordered 95% credible intervals of store-level each store

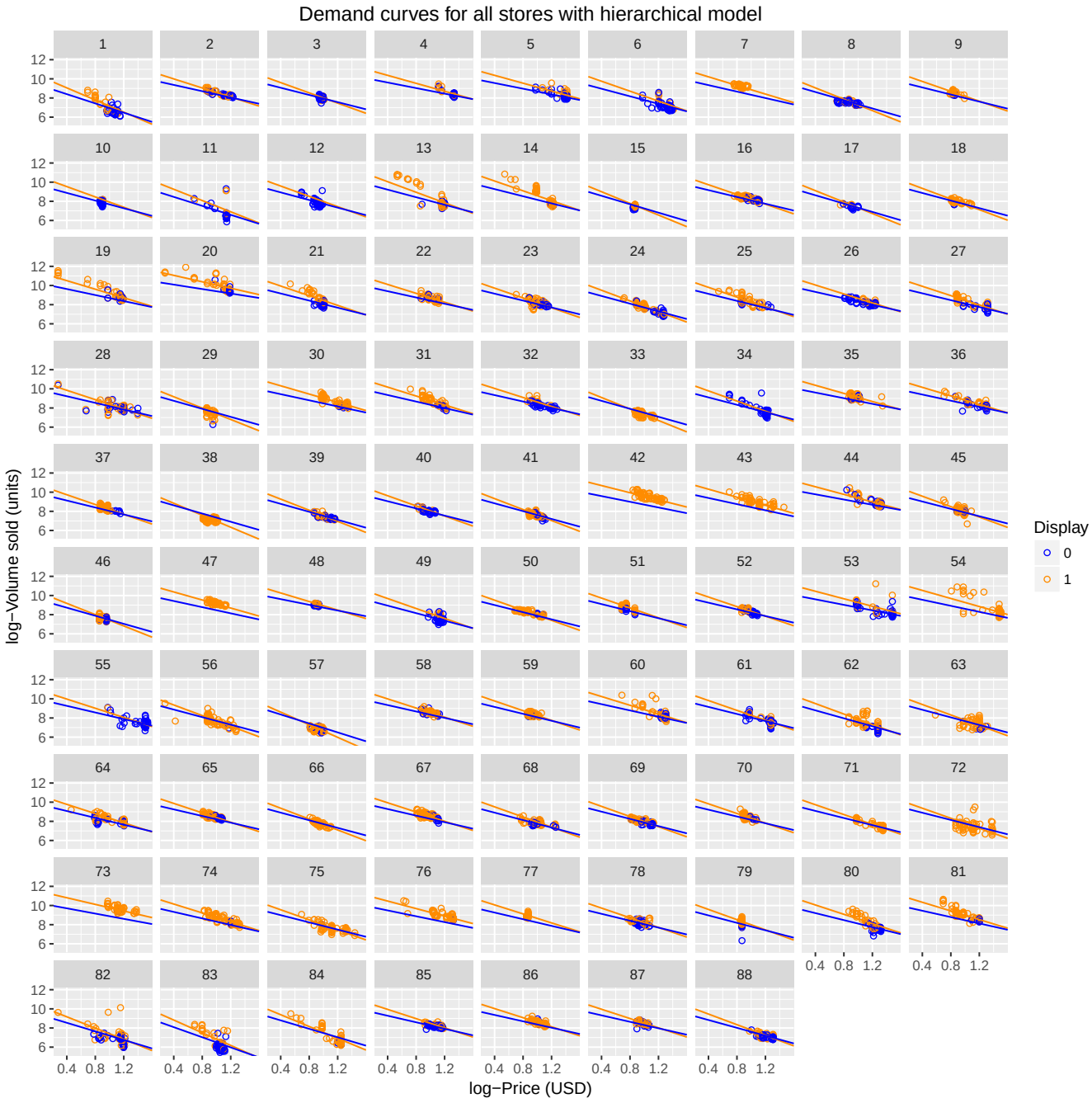


Figure 9: Each store's demand curves with fitted line from Bayesian hierarchical model

# Problem 3

## A hierarchical probit model via data augmentation

For this model we model $y_{ij}$, the $j$th binary 0-1 response, $j \in \{1, 2, \ldots, n_i\}$, within group $i \in \{1, 2, \ldots, I\}$ through the utilization of data augmentation whereby we introduce a latent variable $z_{ij}$,

$$\begin{aligned}(z_{ij}|\beta, \gamma_i) &\sim N(x_{ij}^T\beta + w_{ij}^T\gamma_i, 1) \\ y_{ij} = \mathbf{1}(z_{ij} > 0) &= \begin{cases} 1 & \text{if } z_{ij} > 0 \\ 0 & \text{if } z_{ij} \leq 0 \end{cases},\end{aligned}$$

where $x_{ij}$ is a vector of covariate features and $w_{ij}$ is a subset of these features whose effects vary at the subject level, captured through $\gamma_i$. We can see that this implies a probit link function so that

$$p_{ij} = P(y_{ij} = 1) = \Phi(x_{ij}^T\beta + w_{ij}^T\gamma_i),$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. Let $z_i$ be the $n_i$-length vector of responses from subject $i$, and similarly, $X_i$ is a $n_i \times p$ design matrix of subject $i$ and $W_i$ is a $n_i \times q$ design matrix with $q \leq p$ whose columns are a a subset of the columns of $X_i$. We then see that

$$(z_i|\beta, \gamma_i) \sim \mathcal{N}_{n_i}\left(X_i\beta + W_i\gamma_i, \mathcal{I}_{n_i}\right),$$

and furthermore we model the subject-level responses as coming from a multivatiate normal distribution

$$\gamma_i \overset{\text{iid}}{\sim} \mathcal{N}_q(0, D),$$

where $D$ is some $q \times q$ covariance matrix. We set the priors for our parameters,

$$\begin{aligned}\pi(\beta) &\propto 1 \\ \pi(D) &\sim \text{IW}\left(\nu, \Psi\right),\end{aligned}$$

and now we can show the full conditionals for the Gibbs sampler. At each iteraton we also need to generate values for the latent variables $z_i$.

$$\pi(\Sigma|\gamma_i) \sim \text{IW}\left(\nu + I, \Psi + \sum_{i=1}^{I} \gamma_i\gamma_i^T\right)$$

$$\begin{aligned}\pi\left(\gamma_i|z_i, \beta, D\right) &\propto \pi\left(\gamma_i|D\right) p\left(z_i\right) \\ &\sim \mathcal{N}\left(\left[W_i^TW_i + D^{-1}\right]^{-1} W_i^T(z_i - X_i\beta), \left[W_i^TW_i + D^{-1}\right]^{-1}\right)\end{aligned}$$

$$\begin{aligned}\pi(\beta|z, \gamma) &\propto \pi(\beta) \prod_{i=1}^{I} p\left(z_i|\beta, \gamma_i\right) \\ &\propto \prod_{i=1}^{I} \exp\left[-\frac{1}{2}(z_i - X_i\beta - W_i\gamma_i)^T(z_i - X_i\beta - W_i\gamma_i)\right] \\ &\sim \mathcal{N}\left(\left[\sum_{i=1}^{I} X_i^TX_i\right]^{-1} \sum_{i=1}^{I} X_i^T(z_i - W_i\gamma_i), \left[\sum_{i=1}^{I} X_i^TX_i\right]^{-1}\right)\end{aligned}$$

Finally, the latent variables are generated as follows:

(1)

$$\tilde{z}_i \sim \mathcal{N}_{n_i}(X_i\beta + W_i b_i, I_{n_i})$$

(2) For each $z_{ij}$,

$$z_{ij}|y_{ij} = \begin{cases} \min\{0, \tilde{z}_{ij}\} & \text{if } y_{ij} = 1 \\ \max\{0, \tilde{z}_{ij}\} & \text{if } y_{ij} = 0 \end{cases}$$

# Problem 4

## Gene expression over time

For this problem, we have measurements of the gene-expression profiles of 14 genes in the *Drosophila* genome tracked over time during embryogenesis. Figure 10 shows the data, faceted by each gene. There are two levels of hierarchy in the data, as demonstrated. Each gene belongs to a cluster, or "group" as it is called in this specific context, and each gene has three biological replicates. Figure 11 demonstrates this two-level hierarchical structure; the left column shows the expression profiles for all the genes in each group, and the right column shows the replicates of each gene for a given group. To accomodate the hierarchical and nonlinear time series nature of the data, we introduce a Bayesian hierarchical non-parametric model.

Let $i$ be the subscript for clusters of genes, $n$ in the subscript genes, $r$ is the subscript for replicates. If gene $n$ belongs to cluster $i$ we denote this as $n \in c_i$, and $N_i = \#\{n \in c_i\}$. Each gene $n$ has $N_n$ replicates, and each replicate $r$ of gene $n$ has $N_{nr}$ measuremeants across time. Note that because every array of genes is measured all at once, so each gene has the same $D = \sum_{r=1}^{N_n} N_{nr}$ total measurements across time for all replicates.

We can say, for every replicate $r$ of gene $n$, the data we observe take the form of $\mathbf{y}_{nr}$, a $N_{nr} \times 1$ vector observed at times $\mathbf{t}_{nr}$. Define the following Gaussian processes,

$$\begin{aligned} h_i(t) &\sim \text{GP}(\mathbf{0}, k_h(t,t')) \\ g_n(t) &\sim \text{GP}(h_i(t), k_g(t,t')) \text{ for } n \in c_i \\ f_{nr}(t) &\sim \text{GP}(g_n(t), k_f(t,t')) \end{aligned}$$

for some covariance functions $k_h(t,t')$, $k_g(t,t')$, and $k_f(t,t')$. Suppose we have $\mathbf{h}_i$, $\mathbf{g}_n$, and $\mathbf{f}_{nr}$ which is draws from $h_i(t)$, $g_n(t)$, and $f_{nr}(t)$, respectively, at times $\mathbf{t}_{nr}$. Define $\mathbf{K}_f(\mathbf{t}_{nr}, \mathbf{t}_{nr'})$ to be the $N_{nr} \times N_{nr'}$ matrix such that its $(i,j)$ element is $k_g(\mathbf{t}_{nr}[i], \mathbf{t}_{nr'}[j])$, and define the matrices $\mathbf{K}_g(\mathbf{t}_{nr}, \mathbf{t}_{nr'})$ and $\mathbf{K}_h(\mathbf{t}_{nr}, \mathbf{t}_{nr'})$ likewise. Then we model the data $\mathbf{y}_{nr}$ as

$$\mathbf{y}_{nr} = \mathbf{f}_{nr} + \mathbf{e}, \quad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathcal{I}).$$

We can see the following conditional distributions,

$$\begin{aligned} (\mathbf{y}_{nr} | \mathbf{f}_{nr}) &\sim \mathcal{N}\left(\mathbf{f}_{nr}, \sigma^2 \mathcal{I}\right) \\ (\mathbf{f}_{nr} | \mathbf{g}_n) &\sim \mathcal{N}\left(\mathbf{g}_n, \mathbf{K}_f(\mathbf{t}_{nr}, \mathbf{t}_{nr})\right) \\ (\mathbf{g}_n | \mathbf{h}_i) &\sim \mathcal{N}\left(\mathbf{h}_i, \mathbf{K}_g(\mathbf{t}_{nr}, \mathbf{t}_{nr})\right) \\ (\mathbf{h}_i | \mathbf{t}_{nr}) &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}_h(\mathbf{t}_{nr}, \mathbf{t}_{nr})\right) \end{aligned}$$

It is straightforward to find the marginal likelihood of the data $\mathbf{y}_{nr}$,

$$(\mathbf{y}_{nr} | \mathbf{t}_{nr}, \theta) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}_h(\mathbf{t}_{nr}, \mathbf{t}_{nr}) + \mathbf{K}_g(\mathbf{t}_{nr}, \mathbf{t}_{nr}) + \mathbf{K}_f(\mathbf{t}_{nr}, \mathbf{t}_{nr}) + \sigma^2 \mathcal{I}\right)$$

where $\theta$ is a vector which includes the parameters to all of the covariance functions. Now we consider the full data vector for all genes in cluster $i$, $\mathbf{Y}_i = \{\mathbf{y}_n\}_{n \in c_i}$ where each $\mathbf{y}_n$ is a concatenation of the replicates in gene $n$, $\mathbf{y}_n = \{\mathbf{y}_{nr}\}_{r=1}^{N_n}$, $\mathbf{t}_n = \{\mathbf{t}_{nr}\}_{r=1}^{N_n} =: \mathbf{t}$ is the same for each $n$ as illustrated above, and has length $D$, $\mathbf{T}_i = \{\mathbf{t}_k\}_{k \in c_i}$. This will have a marginal full likelihood,

$$(\mathbf{Y}_i | \mathbf{T}_i, \theta) \sim \mathcal{N}(\mathbf{0}, \Sigma_i),$$

Problem 4 continued on next page... 

where $\Sigma_i$ has a matrix which is $N_i \times N_i$ arrangement of block matrices, each of which is of dimension $D \times D$,

$$\Sigma_i[n, n'] = \begin{cases} \mathbf{K}_h(\mathbf{t}, \mathbf{t}) + \Sigma_n & \text{if } n = n' \\ \mathbf{K}_h(\mathbf{t}, \mathbf{t}) & \text{otherwise} \end{cases},$$

where each $\Sigma_n$ is a covariance matrix representing the with-in gene variance for gene $n$, i.e. the marginal covariance matrix of $\mathbf{y}_n$,

$$\Sigma_n[r, r'] = \begin{cases} \mathbf{K}_g(\mathbf{t}_{nr}, \mathbf{t}_{nr}) + \mathbf{K}_f(\mathbf{t}_{nr}, \mathbf{t}_{nr}) + \sigma^2\mathcal{I} & \text{if } r = r' \\ \mathbf{K}_g(\mathbf{t}_{nr}, \mathbf{t}_{nr'}) & \text{otherwise} \end{cases}$$

and also notice that each block $\Sigma_n[r, r']$ is of dimension $N_{nr} \times N_{nr'}$.

Now suppose we want to find the conditional distribution of "new" draws from the Gaussian grocesses given the data we observe. Specifically we want to find the distribution of $\mathbf{h}_i^\star$ drawn at $\mathbf{t}_i^\star$, $\mathbf{g}_n^\star$ at $\mathbf{t}_n^\star$, and $\mathbf{f}_{nr}^\star$ at $\mathbf{t}_{nr}^\star$, conditional on the data. First, it is easy to find the respective marginal distributions of each of these,

$$\begin{aligned}
(\mathbf{h}_i^\star | \mathbf{t}_i^\star) &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}_h(\mathbf{t}_i^\star, \mathbf{t}_i^\star)\right) \\
(\mathbf{g}_n^\star | \mathbf{t}_n^\star) &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}_h(\mathbf{t}_n^\star, \mathbf{t}_n^\star) + \mathbf{K}_g(\mathbf{t}_n^\star, \mathbf{t}_n^\star)\right) \\
(\mathbf{f}_{nr}^\star | \mathbf{t}_{nr}^\star) &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}_h(\mathbf{t}_{nr}^\star, \mathbf{t}_{nr}^\star) + \mathbf{K}_g(\mathbf{t}_{nr}^\star, \mathbf{t}_{nr}^\star) + \mathbf{K}_f(\mathbf{t}_{nr}^\star, \mathbf{t}_{nr}^\star)\right).
\end{aligned}$$

Conditioned on $\mathbf{y}_i$, the distribution of each becomes

$$\begin{aligned}
\begin{bmatrix} \mathbf{Y}_i \\ \mathbf{h}_i^\star \end{bmatrix} &\sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma_i & \mathbf{K}_{i\star}^T \\ \mathbf{K}_{i\star} & \mathbf{K}_{i\star\star} \end{bmatrix}\right) \\
\begin{bmatrix} \mathbf{Y}_i \\ \mathbf{g}_n^\star \end{bmatrix} &\sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma_i & \mathbf{K}_{n\star}^T \\ \mathbf{K}_{n\star} & \mathbf{K}_{n\star\star} \end{bmatrix}\right) \\
\begin{bmatrix} \mathbf{Y}_i \\ \mathbf{f}_{nr}^\star \end{bmatrix} &\sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma_i & \mathbf{K}_{nr\star}^T \\ \mathbf{K}_{nr\star} & \mathbf{K}_{nr\star\star} \end{bmatrix}\right),
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{K}_{i\star\star} &= \mathbf{K}_h(\mathbf{t}_i^\star, \mathbf{t}_i^\star) \\
\mathbf{K}_{n\star\star} &= \mathbf{K}_h(\mathbf{t}_n^\star, \mathbf{t}_n^\star) + \mathbf{K}_g(\mathbf{t}_n^\star, \mathbf{t}_n^\star) \\
\mathbf{K}_{nr\star\star} &= \mathbf{K}_h(\mathbf{t}_{nr}^\star, \mathbf{t}_{nr}^\star) + \mathbf{K}_g(\mathbf{t}_{nr}^\star, \mathbf{t}_{nr}^\star) + \mathbf{K}_f(\mathbf{t}_{nr}^\star, \mathbf{t}_{nr}^\star)
\end{aligned}$$

and the elements of the off-diagonal matrices are given as

$$\begin{aligned}
\mathbf{K}_{i\star}[t, t'] &= \text{cov}\left(\mathbf{h}_i^\star[t], \mathbf{Y}_i[t']\right) = k_h(t, t') \\
\mathbf{K}_{n\star}[t, t'] &= \text{cov}\left(\mathbf{g}_n^\star[t], \mathbf{Y}_i[t'] \in \mathbf{y}_n\right) = \begin{cases} k_h(t, t') + k_g(t, t') & \text{if } n = n' \\ k_h(t, t') & \text{otherwise} \end{cases} \\
\mathbf{K}_{nr\star}[t, t'] &= \text{cov}\left(\mathbf{f}_{nr}^\star[t], \mathbf{Y}_i[t'] \in \mathbf{y}_{n'r'}\right) = \begin{cases} k_h(t, t') + k_g(t, t') + k_f(t, t') & \text{if } n = n' \text{ and } r = r' \\ k_h(t, t') + k_g(t, t') & \text{if } n = n' \text{ and } r \neq r' \\ k_h(t, t') & \text{otherwise.} \end{cases}
\end{aligned}$$

With all this in hand, the conditional distributions may be written explicitly, e.g.

$$(\mathbf{h}_i^\star | \mathbf{Y}_i) \sim \mathcal{N}\left(\mathbf{K}_{i\star}\Sigma_i^{-1}\mathbf{Y}_i, \mathbf{K}_{i\star\star} - \mathbf{K}_{i\star}\Sigma_i^{-1}\mathbf{K}_{i\star}^T\right).$$

Problem 4 continued on next page...

The challenge now is to choose the hyperparameters within $\theta$ by maximizing the marginal likelihood of the full data vector, $\mathbf{Y}_i$. In my R script, I used the `optim` command to do this, which by default uses the Nelder-Mead method of optimization. For this application, we use the squared-exponential with zero "nugget" parameter, e.g., for the cluster-level Gaussian process,

$$k_h(t, t') = \alpha_h \cdot \exp\left[-\frac{(t-t')^2}{\gamma_h}\right].$$

Results of the HGP regression are shown below, at the group, gene, and replicate level for all three groups. The estimated time series functions are shown, along with a 95% confidence band. Not that we have strange results for Group 1. The estimated optimal $\gamma_h$ parameter obtained by `optim` for Group 1 is very high, around $e^{13}$ This is likely due to a very flat likelihood function, due to the fact that the data all look very similar to each other. The high value of $\gamma_h$ gives a matrix $\mathbf{K}_h(\mathbf{t}, \mathbf{t})$ which has very large numbers in most of its elements, which may have been a source of numerical instability in R.

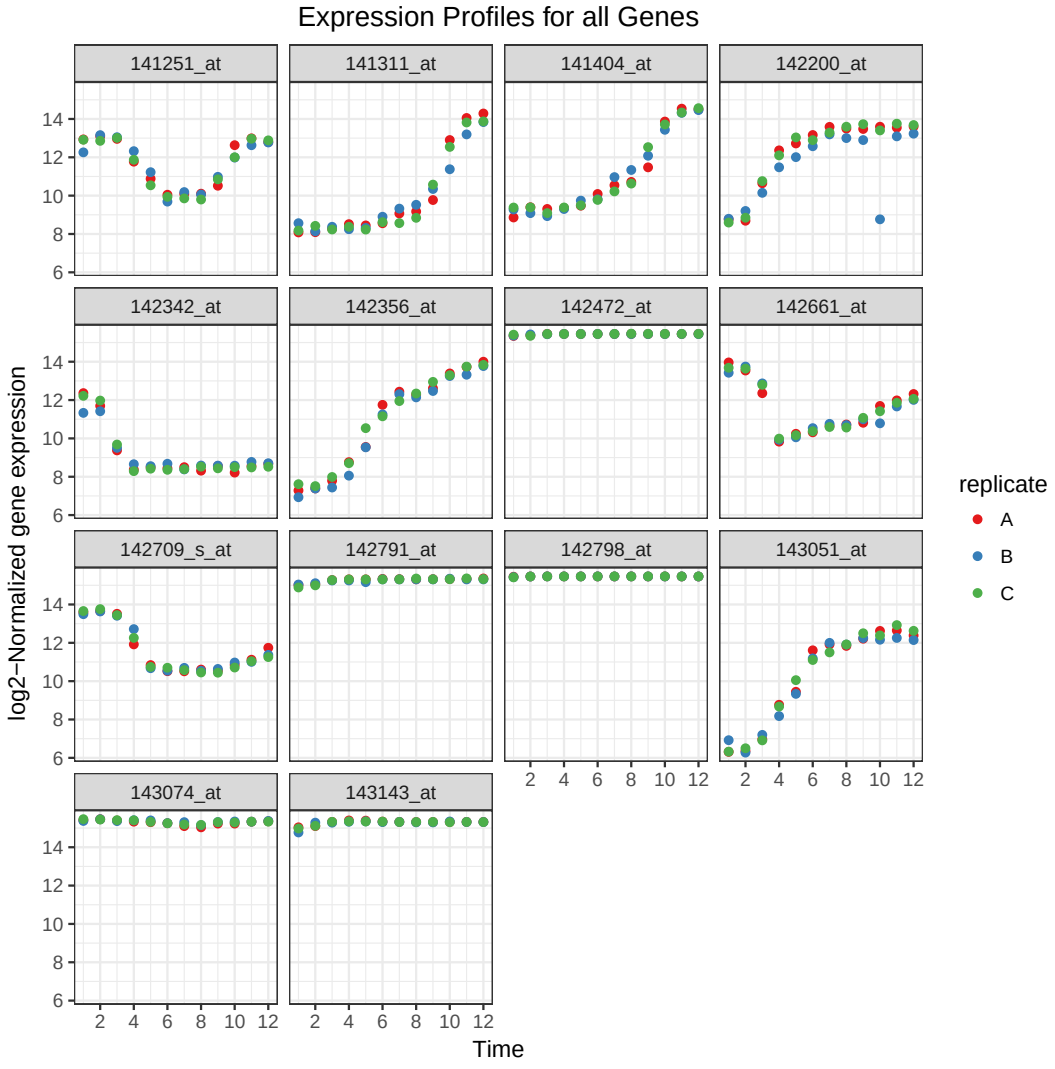


Figure 10: Expression profiles for each gene across all replicates

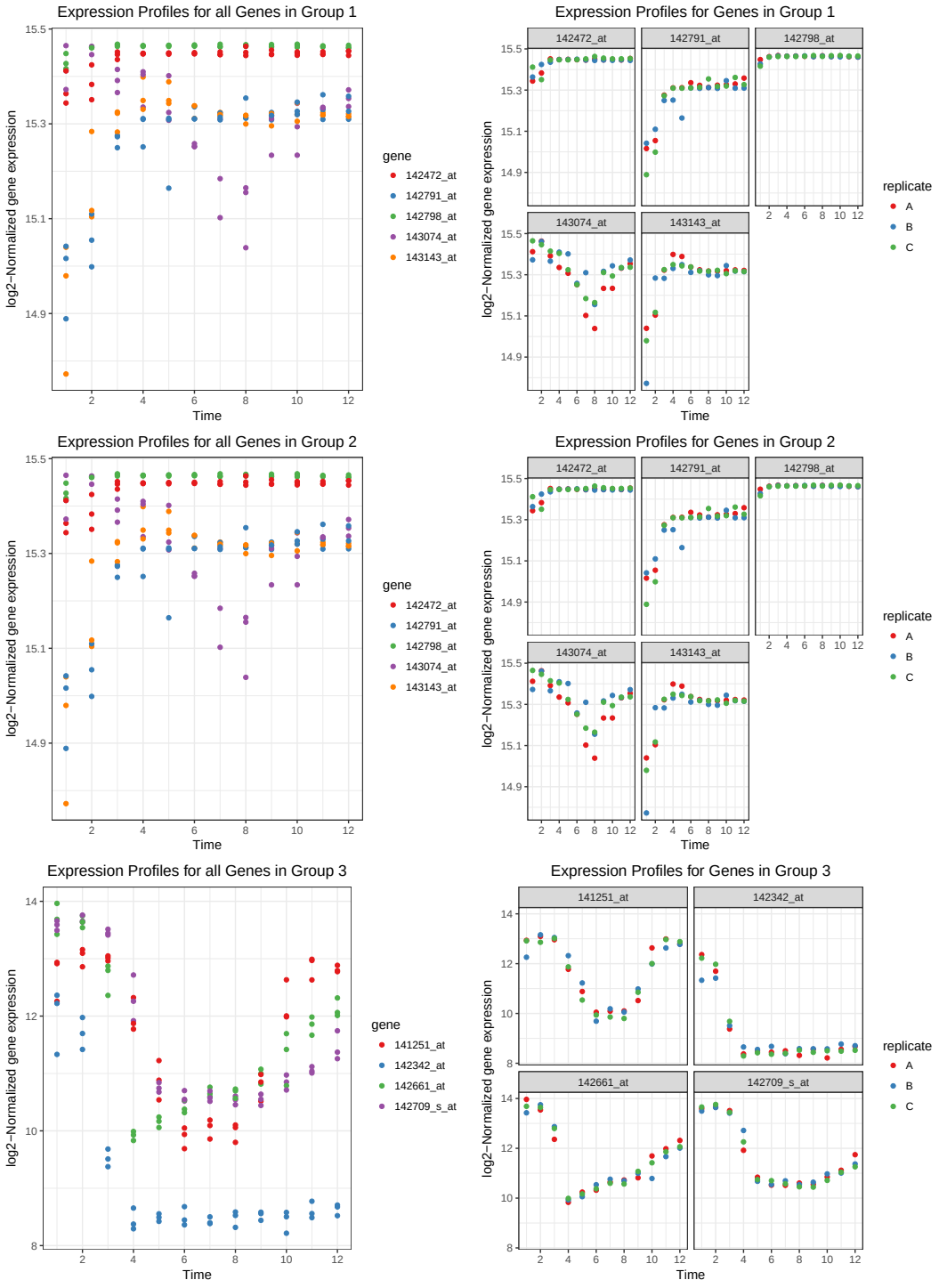


Figure 11: Expression profiles of all genes, accounting for clusters (or "groups") and replicates

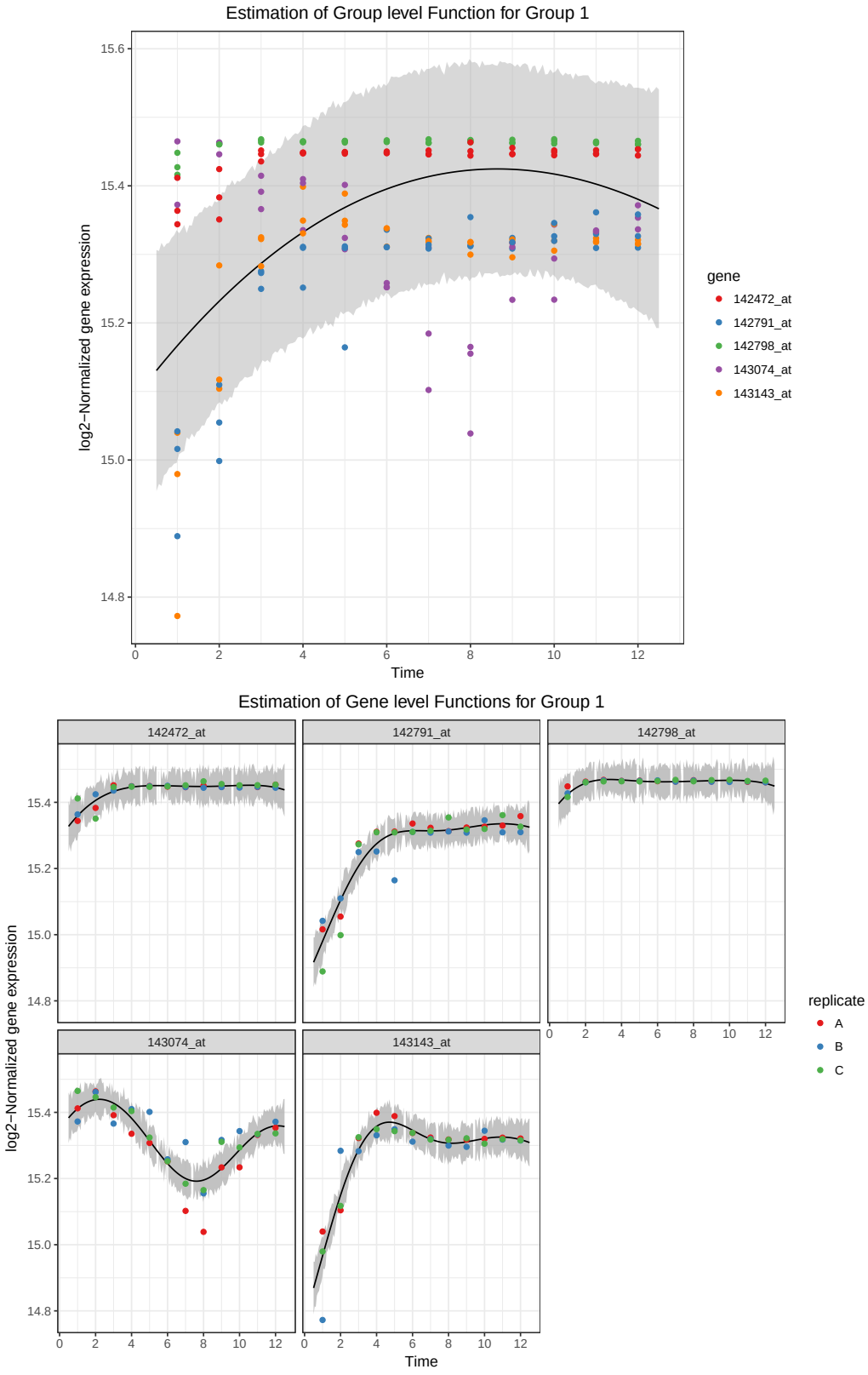


Figure 12: Estimation of group- and gene-level gene expression time series functions for Group 1

Problem 4 continued on next page. . .

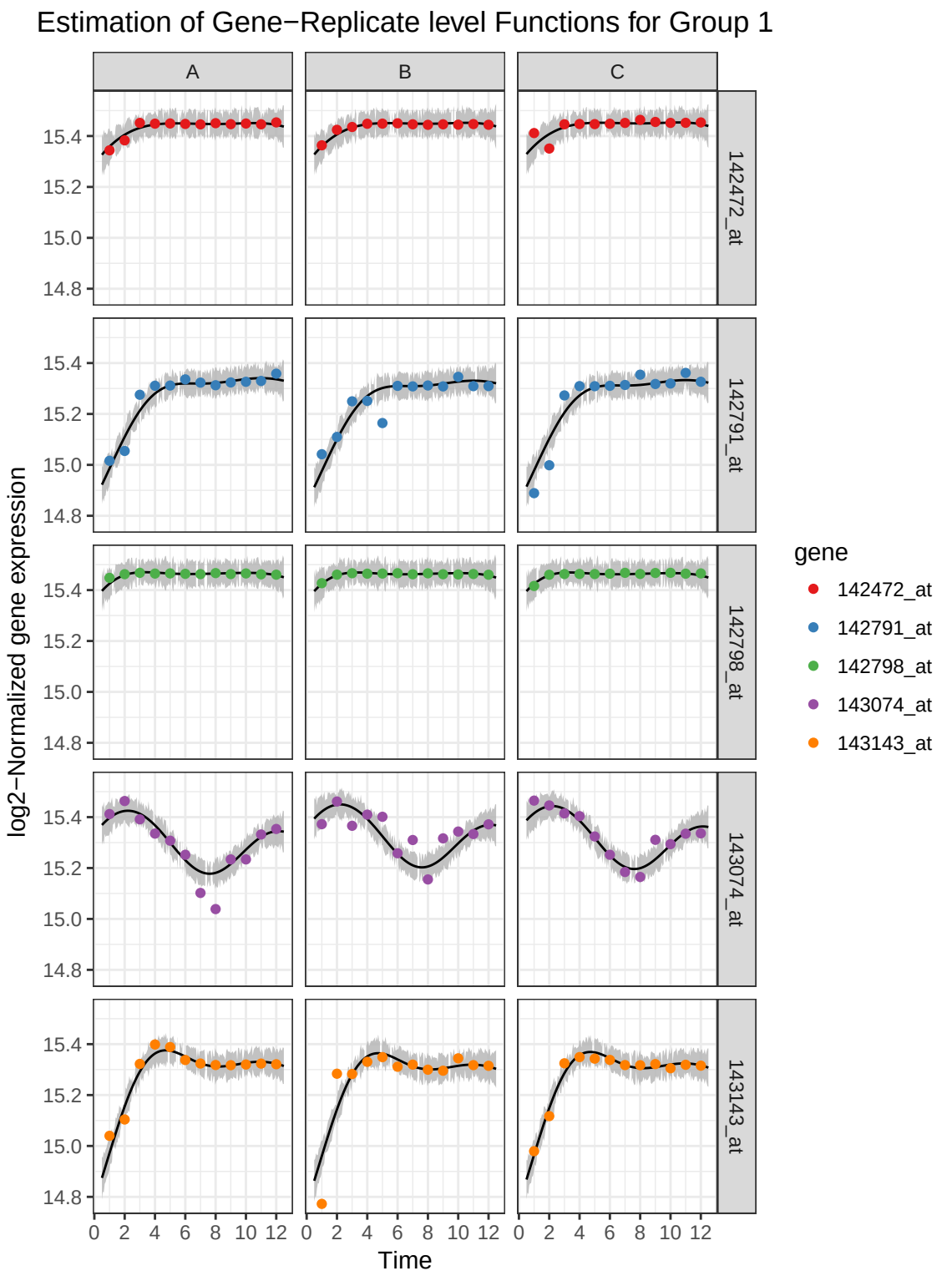


Figure 13: Estimation of gene, replicate-level gene expression time series functions for genes in Group 1

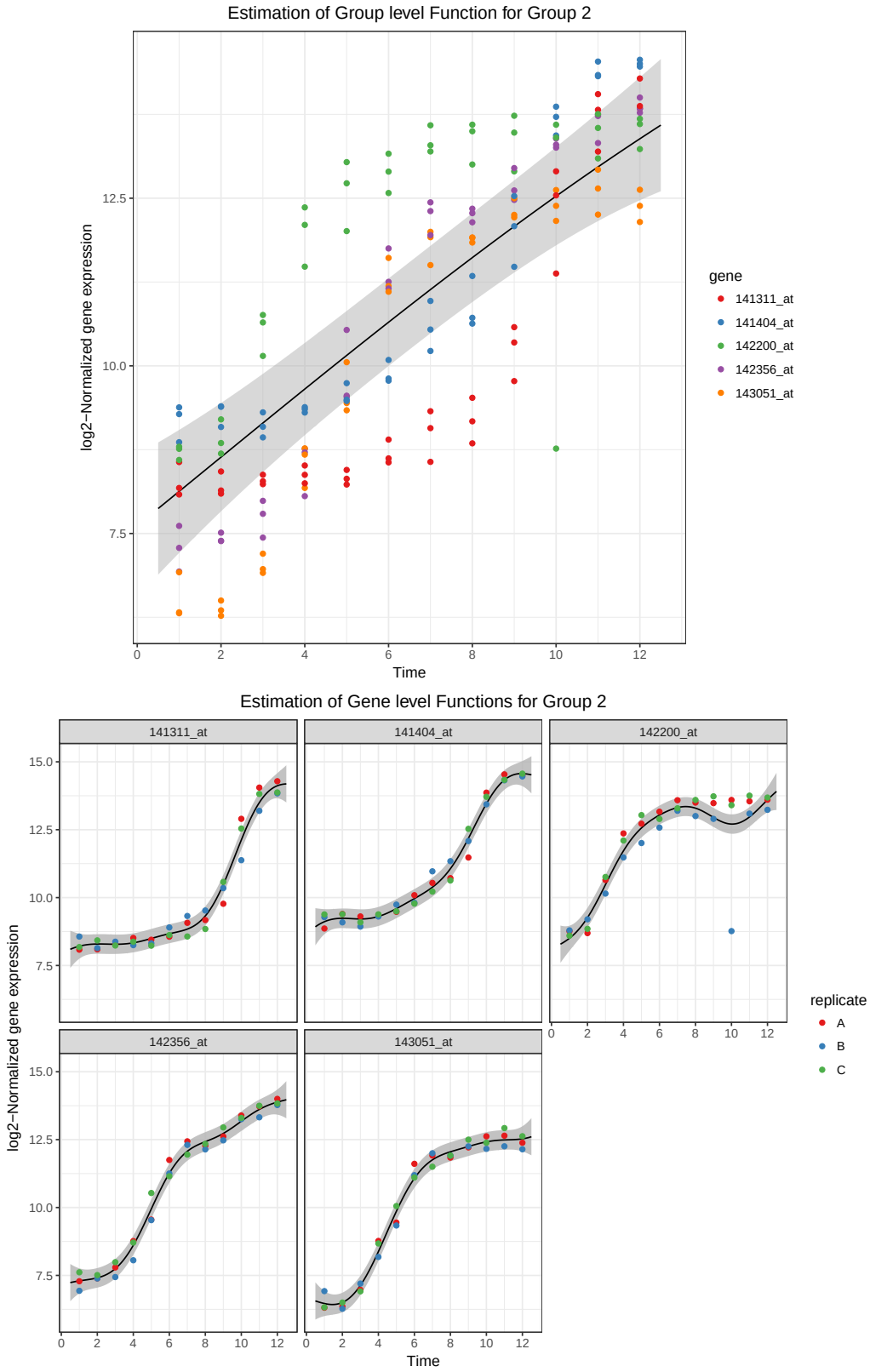


Figure 14: Estimation of group- and gene-level gene expression time series functions for Group 2

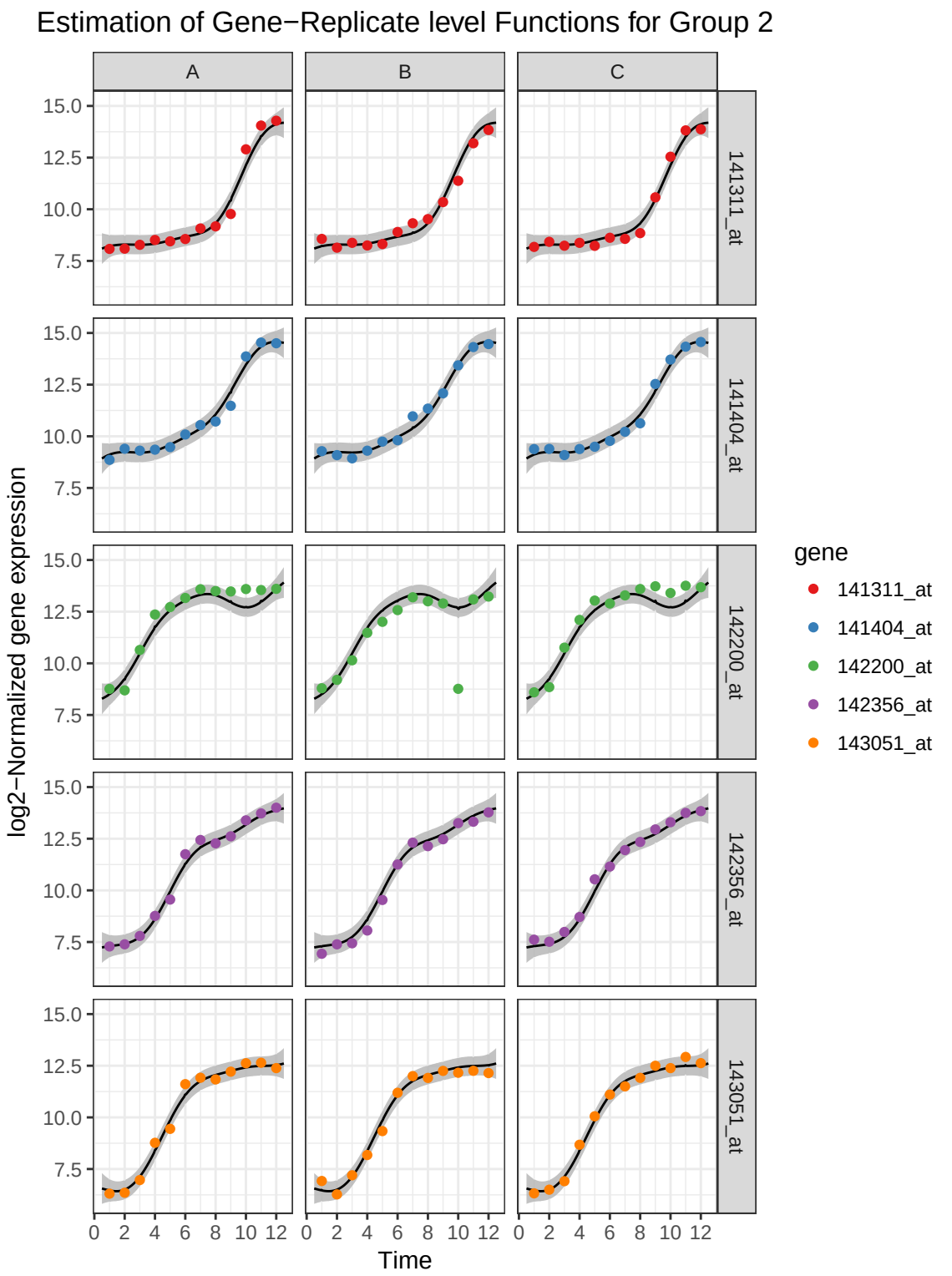


Figure 15: Estimation of gene, replicate-level gene expression time series functions for genes in Group 2

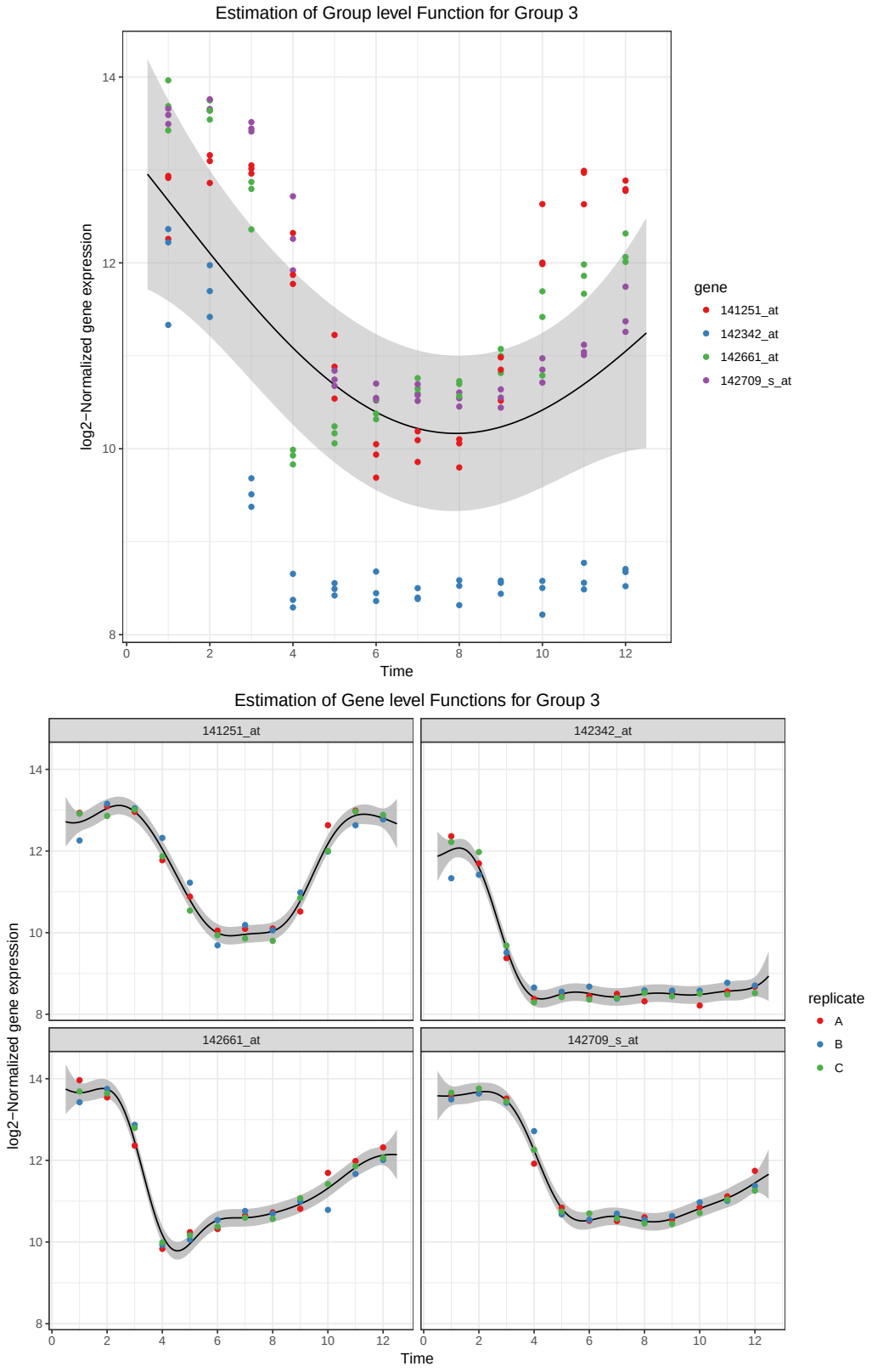


Figure 16: Estimation of group- and gene-level gene expression time series functions for Group 3

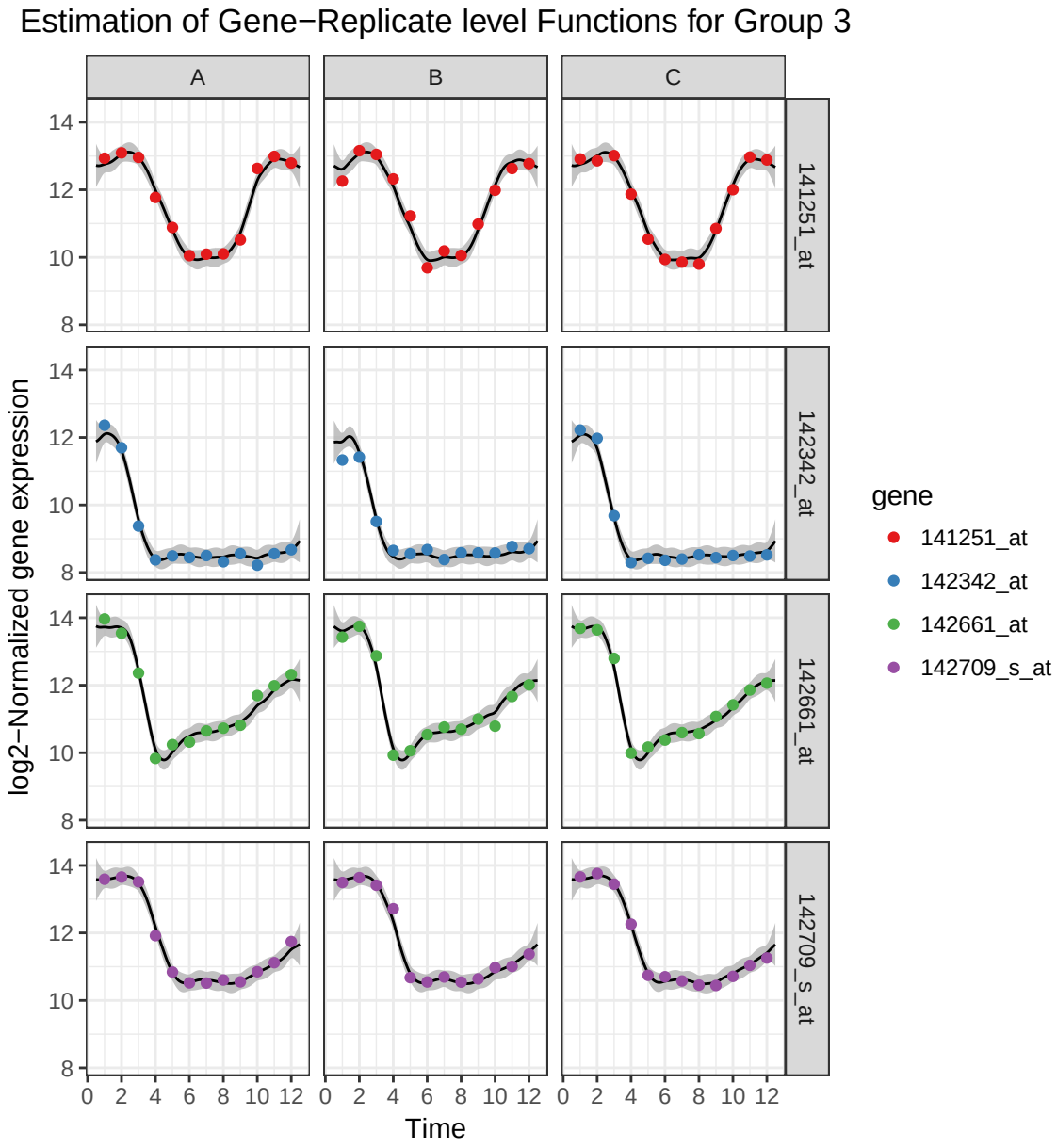


Figure 17: Estimation of gene, replicate-level gene expression time series functions for genes in Group 3