Supplemental Material for "A Resample-Replace Lasso Procedure for Combining High-dimensional Biomarkers with Limit of Detection"

Jinjuan Wang[1], Yunpeng Zhao[2,*], Larry L Tang[3], Claudius Mueller[4], Qizhai Li[5]

[1]Beijing Institute of Technology

[2]Arizona State University

[3]University of Central Florida

[4]George Mason University

[5]LSC, NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

*Corresponding author, E-mail: Yunpeng.Zhao@asu.edu

1 Choice for the Hyperparameter

The proposed method (RIG) applies the linear regression model to the imputation procedure. When the number of the predictor variables is larger than the sample size, it employs principle component analysis and selects the top ranked principal components that explains 80% variance. Here 80% is a hyper-parameter (denoted as q) and may affect the performance of RIG. To investigate the sensitivity of RIG to the choice of q, we set q to be $q_1 = 70\%, q_2 = 80\%$ and $q_3 = 90\%$, respectively, and compare their MSEs in estimating $\{\text{pAUC}(\alpha), \alpha = 0.05, 0.1, 0.2, 1\}$. The MSEs corresponding to $q_1 = 70\%$ are considered as the baseline, and the results are presented in the ratio form as q_2/q_1 and q_3/q_1, indicating those calculated by the division between MSEs of the corresponding q's. The sample size is 50, and the remaining parameters are the same as those stated in the main text.

The results are presented in Table S1. It shows that commonly used threshold values in PCA, i.e., 70%, 80%, 90%, have little difference regarding the performance of RIG. Thus we

Table S1: Ratios of MSE in pAUC(α) estimation for different hyper-parameter.

ratio	ρ	LOD \ α	0.05	0.1	0.15	0.2	1
q_2/q_1	0.2	7.5	1.00000	1.00000	1.00000	1.00000	1.00000
		8.5	1.00003	1.00002	1.00002	1.00002	1.00002
	0.8	7.5	1.00037	1.00037	1.00035	1.00031	1.00033
		8.5	1.01449	1.01400	1.01385	1.01364	1.01227
q_3/q_1	0.2	7.5	1.00000	1.00000	1.00000	1.00000	1.00000
		8.5	1.00012	1.00010	1.00009	1.00008	1.00008
	0.8	7.5	1.00030	1.00028	1.00026	1.00024	1.00023
		8.5	1.01666	1.01616	1.01588	1.01564	1.01409

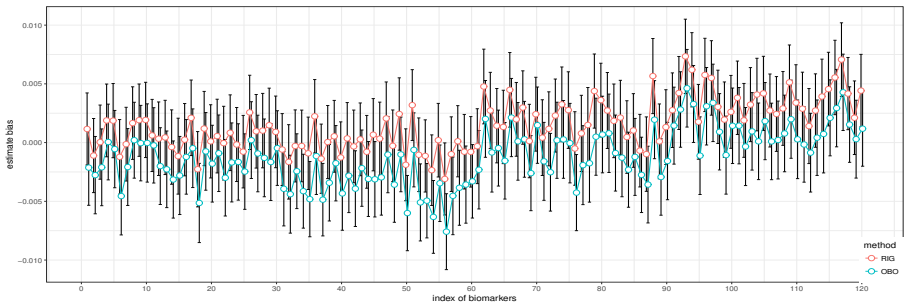
select the top ranked components that explain 80% variance, a common choice in practice.

2 Bias in Parameter Estimation

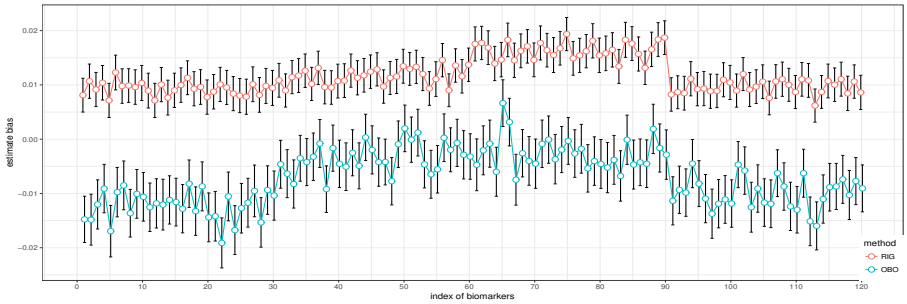
We apply three methods: the resample-input-graphical lasso (RIG), the substituting NA with LOD value method (SNL) and ignoring the observations with NA values (IGN), to the estimation of means and covariance matrices and compare their performance. We set $(m, n) = (100, 100), r = 0.8, d = 8.5$ and keep all the other parameters the same as that described in the Numerical Studies section. Before comparing these three methods directly, we first study the usefulness of taking the correlations between different biomarkers into account. To do this, we estimate the mean and variance of each biomarker via two methods, RIG, the proposed method that uses the correlation information, and OBO, a method that conducts the estimation for each biomarker one by one separately. For each method, we calculate the estimate biases and their corresponding estimate standard errors (SE). The results are shown in Figure S1. The figure shows that when estimating mean, these two

methods have similar absolute bias, and when estimating variance, RIG performs better when the distance between LOD and mean is smaller. In both scenarios, RIG always has smaller SE. We conclude that our method performs no worse than the classical methods. To compare the estimation accuracy of RIG, SNL and IGN, we did the similar calculation and present the results in Figure S2 and S3. Figure S2 shows that for both mean and variance, RIG has smaller estimate bias than SNL, with the estimate standard errors being similar. Comparing (a) and (b)((c) and (d)), we find that the advantage of RIG gets bigger when the distance between population mean and LOD becomes smaller, that is, when more values are missing. This is because under this condition, considering the relationships between multiple variates balances the effect of higher LOD to some extent and makes the superiority of RIG more obvious. Figure S3 indicates similar results.

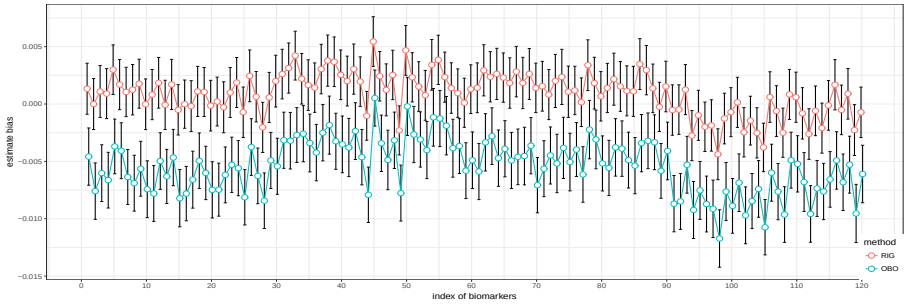
In short, compared with SNL and IGN, RIG is more accurate for estimating the parameters in most of the scenarios considered here.



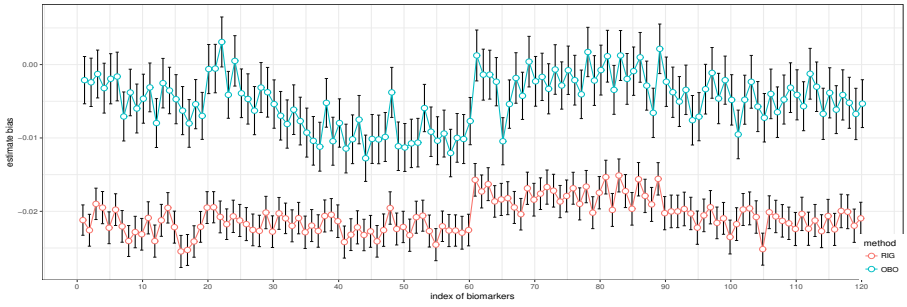
(a)



(b)

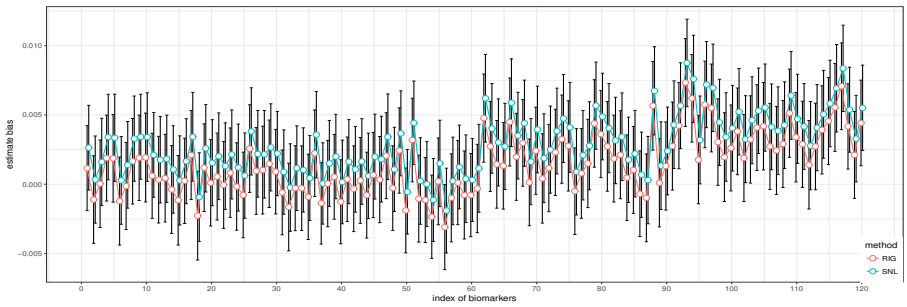


(c)

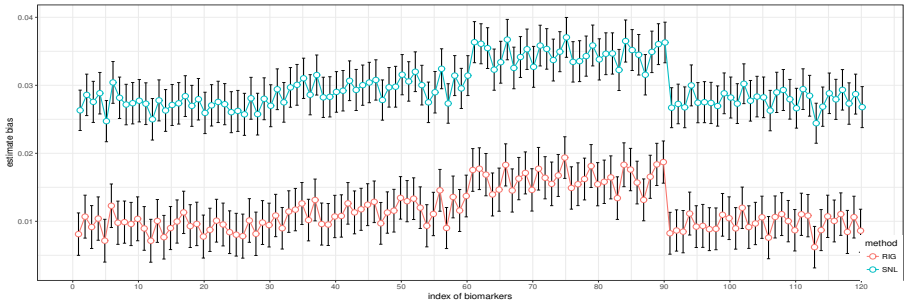


(d)

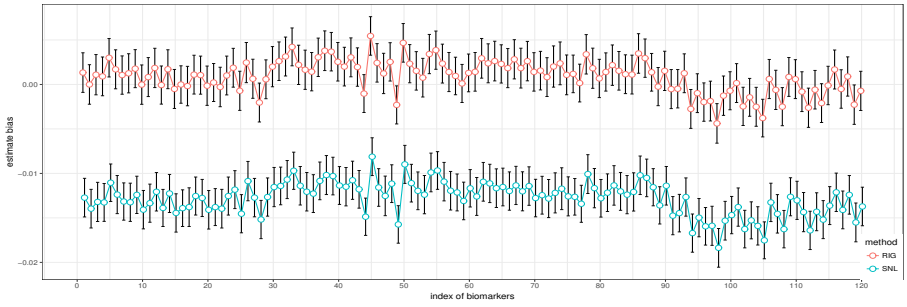
Figure S1: Line chart for the bias in estimation of RIG and OBO. The bars represent standard error(SE). (a) bias for μ; (b) bias for ψ; (c) bias for variance in V; (d) bias for variance in W.



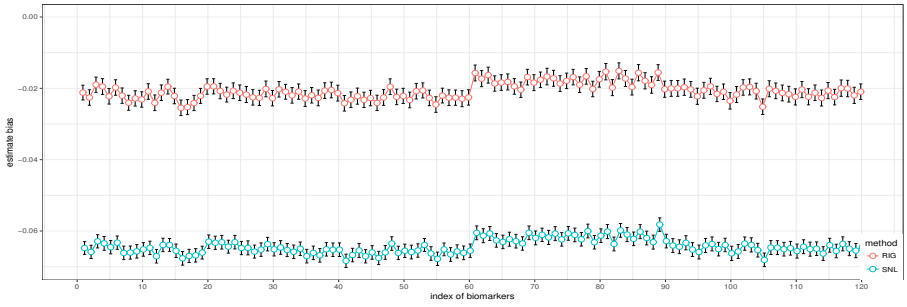
(a)



(b)

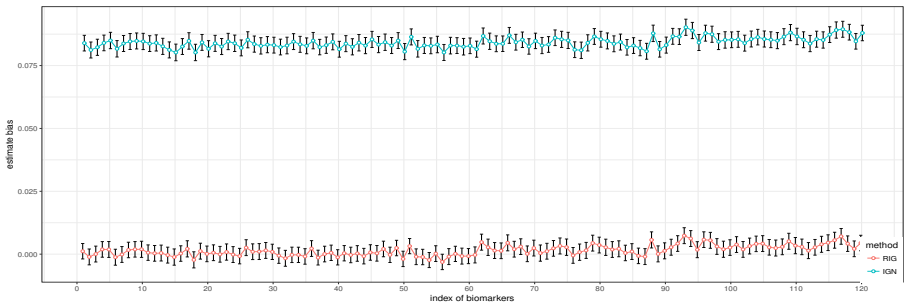


(c)

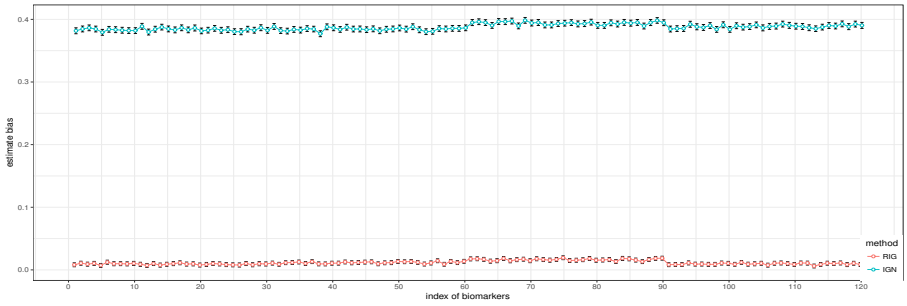


(d)

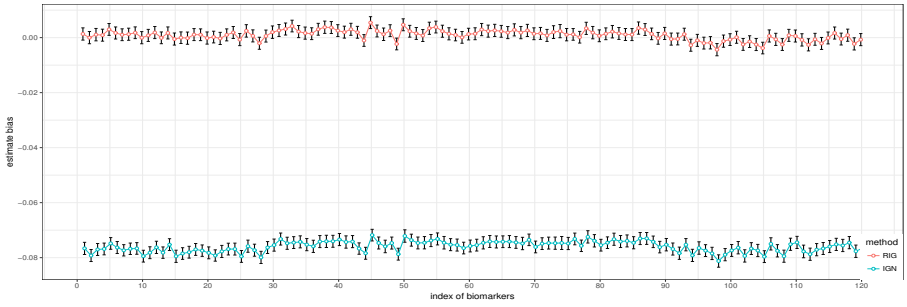
Figure S2: Line chart for estimate bias, using RIG and SNL. The bars represent standard error(SE). (a) mean bias for μ; (b) mean bias for ψ; (c) variance bias for V; (d)variance bias for W.



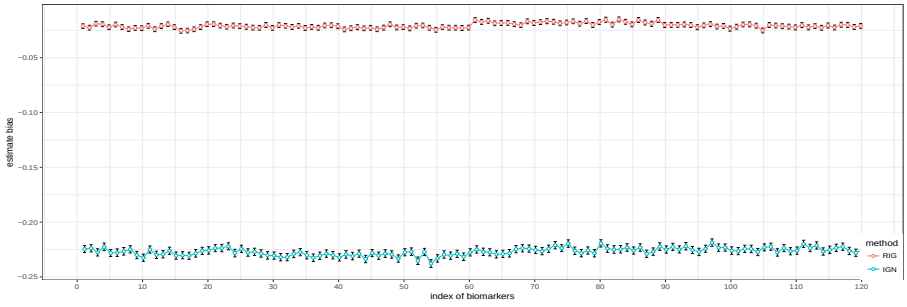
(a)



(b)



(c)



(d)

Figure S3: Line chart for estimate bias, using RIG and IGN. The bars represent standard error(SE). (a) mean bias for μ; (b) mean bias for ψ; (c) variance bias for V; (d)variance bias for W.

3 Asymptotic Results for Imputed Data

We prove that asymptotically the joint distribution of an imputed entry and the observed entries is the same as an observation from the original corresponding unknown population. More precisely, under the multivariate normal assumption, we show that each imputed entry, $\hat{x}_{ik}$, asymptotically follows the same normal distribution $\mathbf{N}(\mu_k, v_{kk})$ as the true latent x_{ik}, and that its covariance with other elements such as x_{il} approximates the true corresponding covariance v_{kl}, as the sample size goes to infinity. Specifically, we prove the following proposition:

Proposition A.1. *Assume* $x_{i1}, \ldots, x_{i,k-1}$, $\mathbf{y}_k = (x_{ik})^\top_{(m-m_k)\times 1}$, $\mathbf{X}_k = (x_{il})_{(m-m_k)\times(k-1)}$ *are observed where* $i = m - m_k + 1, ..., m$ *and* $m - m_k > k$. *Let* x_{ik} *be the true value and* $\hat{x}_{ik}$ *be the imputed value. Then* $(x_{i1}, \ldots, x_{i,k-1}, \hat{x}_{ik}) \xrightarrow{d} (x_{i1}, \ldots, x_{ik})$ *as* $m \to \infty$.

Proof. Our result relies on the correlation between x_{ik} and $x_{i1}, \ldots, x_{i,k-1}$ and thus on the theory of conditional distribution of multivariate normal.

Denote $\omega_k = (\mu_1, \ldots, \mu_k)^\top, \delta_k = (v_{k1}, v_{k2}, \ldots, v_{k,k-1}), \Delta_k = (v_{pq})_{k\times k}, p, q = 1, \ldots, k, k > 1$. For any $i = 1, \ldots, m$, the $k \times 1$ vector $(x_{i1}, \ldots, x_{i,k-1}, x_{ik})$ follows the normal distribution $\mathbf{N}(\omega_k, \Delta_k)$, where

$$\Delta_k = \begin{pmatrix} \Delta_{k-1} & \delta_k^\top \\ \delta_k & v_{kk} \end{pmatrix}.$$

Given $z_{i,k-1} = (x_{i1}, \ldots, x_{i,k-1})$, the conditional distribution of x_{ik} is $\mathbf{N}(\varrho_{ik}, \sigma_k^2)$, with

$$\varrho_{ik} = \mu_k + \delta_k \Delta_{k-1}^{-1}(z_{i,k-1} - \omega_{k-1}),$$

and

$$\sigma_k^2 = v_{kk} - \delta_k \Delta_{k-1}^{-1} \delta_k^\top.$$

Note that ϱ_{ik} is a linear function of $z_{i,k-1}$. Let $\beta_{k0} = \mu_k - \delta_k \Delta_{k-1}^{-1} \omega_{k-1}$ and β_{kl} be the l-th

entry of $\delta_k \Delta_{k-1}^{-1}$. Denote $\beta_k = (\beta_{k0}, \cdots, \beta_{k,k-1})^\top$. Then

$$\varrho_{ik} = \beta_{k0} + \sum_{l=1}^{k-1} \beta_{kl} x_{il}.$$

Equivalently,

$$x_{ik} = \beta_{k0} + \sum_{l=1}^{k-1} \beta_{kl} x_{il} + \epsilon_{ik},$$

where ϵ_{ik} follows $\mathbf{N}(0, \sigma_k^2)$. Denote $\mathbf{Z}_k = [\mathbf{1}_{m-m_k}, \mathbf{X}_k]$, then we can regress $\mathbf{y}_k$ on $\mathbf{Z}_k$ to estimate β_k and σ_k^2. And their corresponding estimates are denoted as $\hat{\beta}_k$ and $\hat{\sigma}_k^2$, respectively. Specifically,

$$\hat{\beta}_k = \left(\hat{\beta}_{k0}, \cdots, \hat{\beta}_{k,k-1}\right)^\top = \left(\mathbf{Z}_k^\top \mathbf{Z}_k\right)^{-1} \mathbf{Z}_k^\top \mathbf{y}_k,$$
$$\hat{\sigma}_k^2 = ||\mathbf{y}_k - \mathbf{Z}_k \hat{\beta}_k||^2/(m - m_k).$$

Then the imputed element is

$$\hat{x}_{ik} = \hat{\beta}_{k0} + \sum_{l=1}^{k-1} \hat{\beta}_{kl} x_{il} + \epsilon_{ik},$$

where ϵ_{ik} is randomly generated from $\mathbf{N}(0, \hat{\sigma}_k^2)$.

Under the theory of least squares estimation, $\mathrm{E}(\hat{\beta}_k) = \beta, \mathrm{var}(\hat{\beta}_k) = v_{kk}(\mathbf{Z}_k^\top \mathbf{Z}_k)^{-1}$. So when $m \to \infty$, each element of $\mathrm{cov}(\hat{\beta}_k)$ is $\mathrm{O}(\frac{1}{m^2})$. That is, $\mathrm{var}(\hat{\beta}_{kl}) = \mathrm{O}(\frac{1}{m^2})$ for $l = 0, 1, \ldots, k-1$, and $\mathrm{cov}(\hat{\beta}_{kl_1}, \hat{\beta}_{kl_2}) = \mathrm{O}(\frac{1}{m^2})$ for $l_1, l_2 = 0, 1, \ldots, k-1, l_1 \neq l_2$. The expectation of $\hat{x}_{ik}$ is

$$\mathrm{E}(\hat{x}_{ik}) = \mathrm{E}(\beta_{k0} + \sum_{l=1}^{k-1} \beta_{kl} x_{il}) + \mathrm{E}(\hat{\beta}_{k0} - \beta_{k0}) + \sum_{l=1}^{k-1} \mathrm{E}((\hat{\beta}_{kl} - \beta_{kl}) x_{il}) + \mathrm{E}(\epsilon_{ik}).$$

The second term equals 0. And $\hat{\beta}_{kl}$ for $l = 1, \ldots, k-1$ is estimated by using $\mathbf{y}_k$ and $\mathbf{Z}_k$, which are independent with x_{il}, so the third term also equals 0. Therefore, $\mathrm{E}(\hat{x}_{ik}) = \mu_k$.

As for the variance of $\hat{x}_{ik}$,

$$\begin{aligned}
\mathrm{var}(\hat{x}_{ik}) &= \mathrm{var}\left(\hat{\beta}_{k0} + \sum_{l=1}^{k-1} \hat{\beta}_{kl} x_{il} + \epsilon_{ik}\right) \\
&= \mathrm{var}\left(\beta_{k0} + \sum_{l=1}^{k-1} \beta_{kl} x_{il}\right) + \mathrm{var}(\hat{\beta}_{k0} - \beta_{k0}) + \mathrm{var}\left(\sum_{l=1}^{k-1} (\hat{\beta}_{kl} - \beta_{kl}) x_{il}\right) + \mathrm{var}(\epsilon_{ik}) \\
&\quad + 2\mathrm{cov}\left(\beta_{k0} + \sum_{l=1}^{k-1} \beta_{kl} x_{il}, \hat{\beta}_{k0} - \beta_{k0}\right) + 2\mathrm{cov}\left(\beta_{k0} + \sum_{l=1}^{k-1} \beta_{kl} x_{il}, \sum_{l=1}^{k-1} (\hat{\beta}_{kl} - \beta_{kl}) x_{il}\right) \\
&\quad + 2\mathrm{cov}\left(\hat{\beta}_{k0} - \beta_{k0}, \sum_{l=1}^{k-1} (\hat{\beta}_{kl} - \beta_{kl}) x_{il}\right) \\
&\quad + 2\mathrm{cov}\left(\beta_{k0} + \sum_{l=1}^{k-1} \beta_{kl} x_{il}, \epsilon_{ik}\right) + 2\mathrm{cov}(\hat{\beta}_{k0} - \beta_{k0}, \epsilon_{ik}) + 2\mathrm{cov}\left(\sum_{l=1}^{k-1} (\hat{\beta}_{kl} - \beta_{kl}) x_{il}, \epsilon_{ik}\right).
\end{aligned}$$

According to $\mathrm{var}(\hat{\beta}_{kl}) = \mathrm{O}(\frac{1}{m^2})$ for $l = 0, 1, \ldots, k-1$ and $\mathrm{cov}(\hat{\beta}_{kl_1}, \hat{\beta}_{kl_2}) = \mathrm{O}(\frac{1}{m^2})$ for $l_1, l_2 = 0, 1, \ldots, k-1, l_1 \neq l_2$, we have that $\mathrm{var}(\hat{\beta}_{k0} - \beta_{k0}) \to 0$, $\mathrm{var}((\hat{\beta}_{kl} - \beta_{kl}) x_{il}) \to 0$, $\mathrm{cov}(\hat{\beta}_{k0} - \beta_{k0}, (\hat{\beta}_{kl} - \beta_{kl}) x_{il}) \to 0$, etc. So in the above equation, except for the first and the forth term, all the other terms asymptotic equals to 0. Thus we have

$$\begin{aligned}
\mathrm{var}(\hat{x}_{ik}) &\to \mathrm{var}\left(\beta_{k0} + \sum_{l=1}^{k-1} \beta_{kl} x_{il}\right) + \sigma_k^2 \\
&= \mathrm{var}(\mu_k + \delta_k \Delta_{k-1}^{-1} (z_{i,k-1} - \omega_{k-1})) + \sigma_k^2 \\
&= \delta_k \Delta_{k-1}^{-1} \Delta_{k-1} \Delta_{k-1}^{-1} \delta_k^\top + \sigma_k^2 \\
&= \delta_k \Delta_{k-1}^{-1} \Delta_{k-1} \Delta_{k-1}^{-1} \delta_k^\top + (v_{kk} - \delta_k \Delta_{k-1}^{-1} \delta_k^\top) \\
&= v_{kk}.
\end{aligned}$$

Similarly,

$$
\begin{aligned}
\operatorname{cov}(z_{i,k-1}, \hat{x}_{ik}) &= \operatorname{cov}\left(z_{i,k-1}, \hat{\beta}_{k0} + \sum_{l=1}^{k-1} \hat{\beta}_{kl} x_{il} + \epsilon_{ik}\right) \\
&= \operatorname{cov}\left(z_{i,k-1}, \beta_{k0} + \sum_{l=1}^{k-1} \beta_{kl} x_{il}\right) + \operatorname{cov}\left(z_{i,k-1}, (\hat{\beta}_{k0} - \beta_{k0}) + \sum_{l=1}^{k-1} (\hat{\beta}_{kl} - \beta_{kl}) x_{il} + \epsilon_{ik}\right) \\
&\to \operatorname{cov}\left(z_{i,k-1}, \beta_{k0} + \sum_{l=1}^{k-1} \beta_{kl} x_{il} + \epsilon_{ik}\right) \\
&= \operatorname{cov}(z_{i,k-1}, \mu_k + \delta_k \Delta_{k-1}^{-1}(z_{i,k-1} - \omega_{k-1})) \\
&= \operatorname{cov}(z_{i,k-1}, \delta_k \Delta_{k-1}^{-1} z_{i,k-1}) \\
&= \Delta_{k-1} \Delta_{k-1}^{-1} \delta_k^\top \\
&= \delta_k^\top .
\end{aligned}
$$

Finally, note that $(x_{i1}, \ldots, x_{i,k-1}, \hat{x}_{ik})$ is asymptotically a linear transformation of normal random variables since $\hat{\beta}_{kl} \to \beta_{kl}$ and $\hat{\sigma}_k^2 \to \sigma_k^2$, and thus asymptotically follows a multivariate normal distribution. Its distribution is determined by the mean and covariance. This completes the proof.