# 检查缩放特征范围、矩阵表示与临时对象

完整操作教程 · 参考固定 · 往返恢复验收

检查ScaleData的真实特征范围、层和矩阵实现，在明确的内存型路径中比较分块并验收输出。

运行边界：示例代码需按安装版本和真实项目输入复核；未提供真实生物数据，语法或小样本检查不代表科学结果和性能验收。

原始counts是稀疏矩阵，并不意味着ScaleData的所有结果都同样稀疏。中心化、缩放或回归后的数据可能有不同表示，所选特征数和细胞数也会直接影响需要存放的结果。遇到内存上升时，先检查实际计算范围，再判断是否真的需要更多容量。

本文讨论常规RNA的NormalizeData、特征选择和ScaleData路径。SCTransform有自己的模型与残差处理，不应机械地在SCT之后重复同一串步骤。当前Seurat还存在面向IterableMatrix的专门实现，内存型矩阵的经验不能不加检查地套到磁盘后端。

## 先列出实际要缩放的特征

查看默认assay、VariableFeatures、data层和矩阵类。某份脚本显式传入全部基因时，它与仅缩放选定变异特征的任务规模不同。如果后续PCA只使用合理的特征集合，首先核对是否有必要创建全特征结果；若研究方法确实要求更多特征，应保留理由，而不是只为了省内存随意删掉目标。

```
library(Seurat)
obj <- readRDS("rna-normalized.rds")
DefaultAssay(obj) <- "RNA"
print(packageVersion("Seurat"))
print(class(obj[["RNA"]]))
print(Layers(obj[["RNA"]]))
ly <- grep("^data($|\.)", Layers(obj[["RNA"]]), value=TRUE)
if (length(ly) != 1L) stop("[redacted]data[redacted]")
m <- LayerData(obj, assay="RNA", layer=ly)
print(list(class=class(m), dim=dim(m)))
features <- VariableFeatures(obj)
stopifnot(length(features)>0L, all(features %in% rownames(m)))
print(length(features))
rm(m)
```

缺少变异特征时，先按照计划的方法完成选择。不要简单把空集合替换成所有基因，让代码继续执行；这可能既改变方法也放大内存。若存在多个data层或磁盘矩阵，应回到对应版本与assay实现检查，而不是为了获取矩阵直接强制JoinLayers或as.matrix。

| 检查点 | 它为何影响占用 | 应采取的判断 |
| --- | --- | --- |
| features | 决定结果包含多少特征 | 与下一步分析要求对齐 |
| 矩阵类和后端 | 决定存储与计算实现 | 不要主动稠密化 |
| vars.to.regress | 增加模型计算与临时数据 | 只按研究设计选择协变量 |
| block.size | 部分实现中影响单次特征块 | 可比较峰值，但不能消除最终输出 |
| 已有scale.data | 旧结果与新结果可能同时存在 | 在独立副本中测试并记录峰值 |

## 在明确的内存矩阵路径上测试范围

下面示例仅适用于已经验收的单层、内存型稀疏data输入。保留原RDS，输出到新文件。块大小是演示设置，应根据当前方法和资源测量调整；它不代表所有ScaleData方法都采用同样的分块机制。

```
m <- LayerData(obj, assay="RNA", layer=ly)
if (!inherits(m, "sparseMatrix")) {
  stop("[redacted]")
}
```

```
rm(m)
scaled <- ScaleData(obj, assay="RNA", features=features,
                    block.size=500, verbose=TRUE)
z <- LayerData(scaled, assay="RNA", layer="scale.data")
stopifnot(all(features %in% rownames(z)))
print(list(class=class(z), dim=dim(z)))
saveRDS(scaled, "rna-scaled-selected-features.rds")
```

## 为什么减小块仍可能失败

降低单次计算块可能减少部分临时占用，但最终结果、原对象、回归数据和其他临时对象仍需存放。它不能将一个方法上必须完整输出的巨大矩阵自动变成流式结果。外层用进程资源工具记录峰值，R的object.size只辅助理解返回对象体积，两者不能混用。

如果更换块大小同时更换特征集合或回归变量，就不能把差异都归因于分块。每次只改变一个因素，保留相同输入和输出要求。避免用更激进的过滤解决技术问题后，又把结果宣称为原任务的等价完成。

## 验收后再决定是否扩容

核对scale.data的特征与细胞集合、矩阵类和后续PCA能否读取所需特征。必要时抽取小范围核查数值是否有限，并重新载入保存文件。不要在大矩阵上额外生成整块逻辑副本，只为做一次没有分块的全量检查。

## 中心化为何可能改变稀疏性

对一个均值大于零的特征，原本为零的条目在减去均值后就可能变成非零值，因此不能从counts里零很多推断缩放后仍然节省同样空间。具体实现可以采用不同存储或延迟计算策略，所以应同时检查数学操作和实际后端，而不是只看输入类名。

零方差特征、非常少的有效观测或协变量问题，还可能影响数值和模型。发现异常时，先按方法核对数据，不以增加内存替代这些判断。数值抽查可以分小范围完成，避免为一次全量检查再生成同尺寸临时对象。

特征范围和实现已经合理、必要计算仍超过可用内存时，可以整理缩放特征数、细胞数、矩阵类、回归项与峰值日志，评估大内存独享CPU租赁。若根因是错误层或方法不兼容，应先修复，再测量资源。

## 参考与版本核对

## 配套练习包

本PDF对应的ZIP包含本文代码块、主题练习模板或合成数据、README和SHA256校验文件。请从同一教程页面下载练习包；先核对校验值，再在独立目录复现。

## 参考资料

Seurat：ScaleData参考 · 来源资料

Seurat：SCTransform参考 · 来源资料

Seurat：BPCells磁盘矩阵与保存 · 来源资料

R官方手册：内存限制与对象 · 来源资料

操作建议：先在合成小样本上验证流程，再依据项目参考、标签合同、存储策略与获配资源验收真实数据。