Supervised Learning: Classification, Part II

Jean Feng & Ali Shojaie

Aug 19-21, 2024
Summer Institute in Statistics for Big Data
University of Washington

Batch Effects

Batch Effects

- In any sort of biological/lab/omics experiment, need to be very aware of batch effects, induced by non-biological factors such as inter-machine or inter-lab or inter-operator variability, time of day, day of week, position of ceiling fan, ...

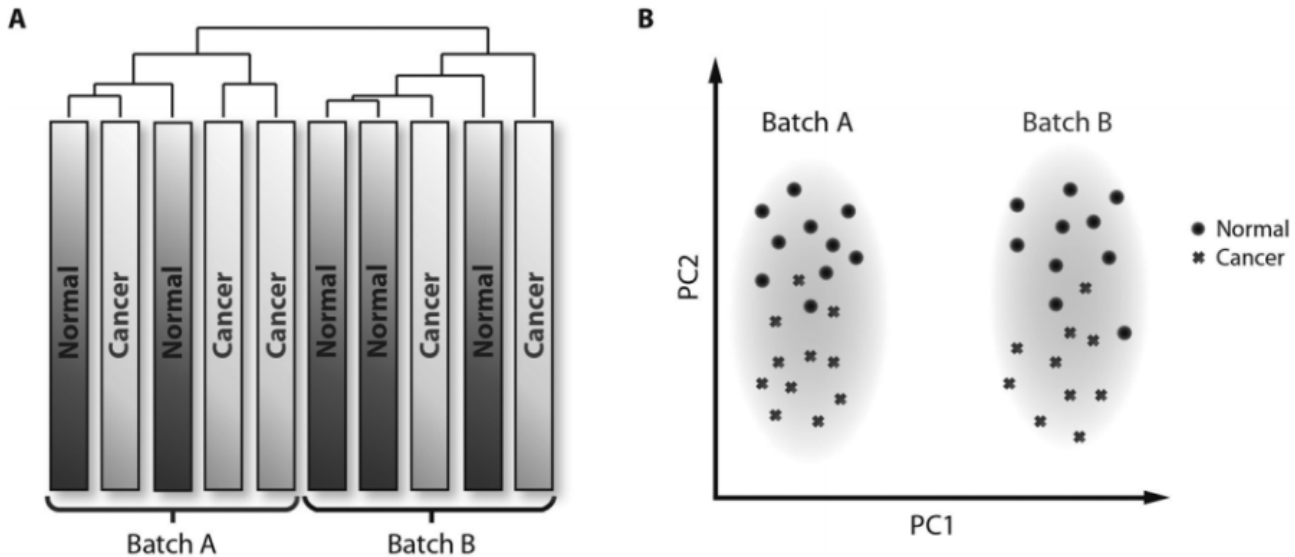
Batch Effects

- In any sort of biological/lab/omics experiment, need to be very aware of batch effects, induced by non-biological factors such as inter-machine or inter-lab or inter-operator variability, time of day, day of week, position of ceiling fan, ...
- It has been shown many many times that batch effects can be much stronger than biological effects of interest!

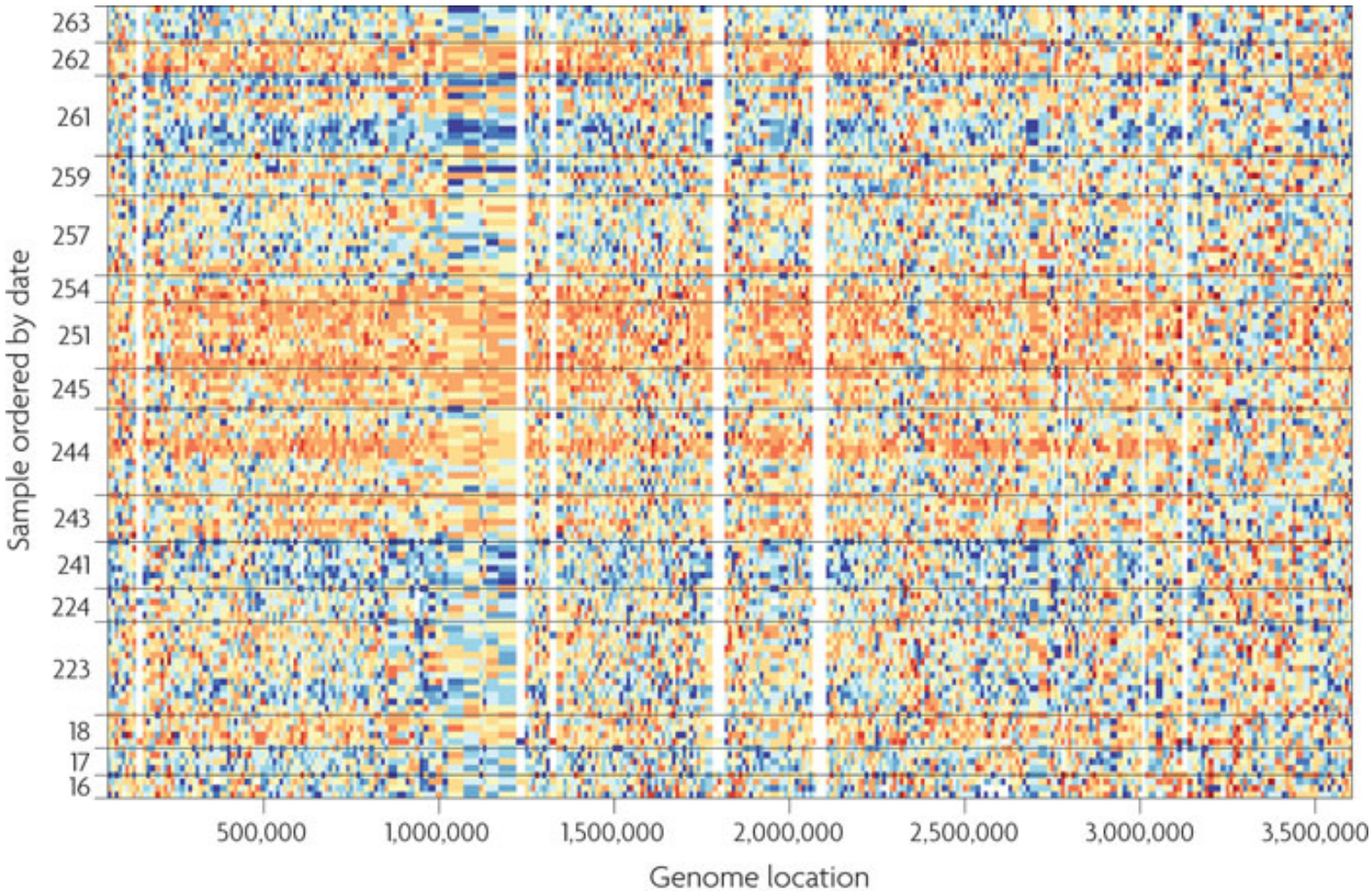
Batch Effects

- In any sort of biological/lab/omics experiment, need to be very aware of batch effects, induced by non-biological factors such as inter-machine or inter-lab or inter-operator variability, time of day, day of week, position of ceiling fan, ...
- It has been shown many many times that batch effects can be much stronger than biological effects of interest!
- Batch effects can make your data nonsense...

Batch Effects



Batch Effects in Practice



Nature Reviews | Genetics

Steps to Reduce Batch Effects

Steps to Reduce Batch Effects

- Randomize sample run times: e.g. don't run cases first and controls second.

Steps to Reduce Batch Effects

- Randomize sample run times: e.g. don't run cases first and controls second.
- Avoid any extraneous sources of variation, e.g. due to change in person running the experiment.

Steps to Reduce Batch Effects

- Randomize sample run times: e.g. don't run cases first and controls second.
- Avoid any extraneous sources of variation, e.g. due to change in person running the experiment.
- It is often better to train a classification or regression method using multiple data sets collected at different institutions, rather than using a single data set.

Steps to Reduce Batch Effects

- Randomize sample run times: e.g. don't run cases first and controls second.
- Avoid any extraneous sources of variation, e.g. due to change in person running the experiment.
- It is often better to train a classification or regression method using multiple data sets collected at different institutions, rather than using a single data set.
- Need to validate any results obtained on independent data sets from a different institution.

Steps to Reduce Batch Effects

- Randomize sample run times: e.g. don't run cases first and controls second.
- Avoid any extraneous sources of variation, e.g. due to change in person running the experiment.
- It is often better to train a classification or regression method using multiple data sets collected at different institutions, rather than using a single data set.
- Need to validate any results obtained on independent data sets from a different institution.

Batch effects are almost inevitable. But you can do your best to design an experiment and analyze the data in such a way that batch effects do not compromise the results obtained.

Subtypes of Breast Cancer

Subtypes of Breast Cancer

- In the past 10 years, global gene expression analyses have identified at least 4 subtypes of breast cancer: Luminal A, Luminal B, Her2-enriched, and basal-like.

Subtypes of Breast Cancer

- In the past 10 years, global gene expression analyses have identified at least 4 subtypes of breast cancer: Luminal A, Luminal B, Her2-enriched, and basal-like.
- Subgroups differ with respect to risk factors, incidence, baseline prognoses, responses to therapies.

Subtypes of Breast Cancer

- In the past 10 years, global gene expression analyses have identified at least 4 subtypes of breast cancer: Luminal A, Luminal B, Her2-enriched, and basal-like.
- Subgroups differ with respect to risk factors, incidence, baseline prognoses, responses to therapies.
- Want to be able to determine the subtype for a new patient with breast cancer.

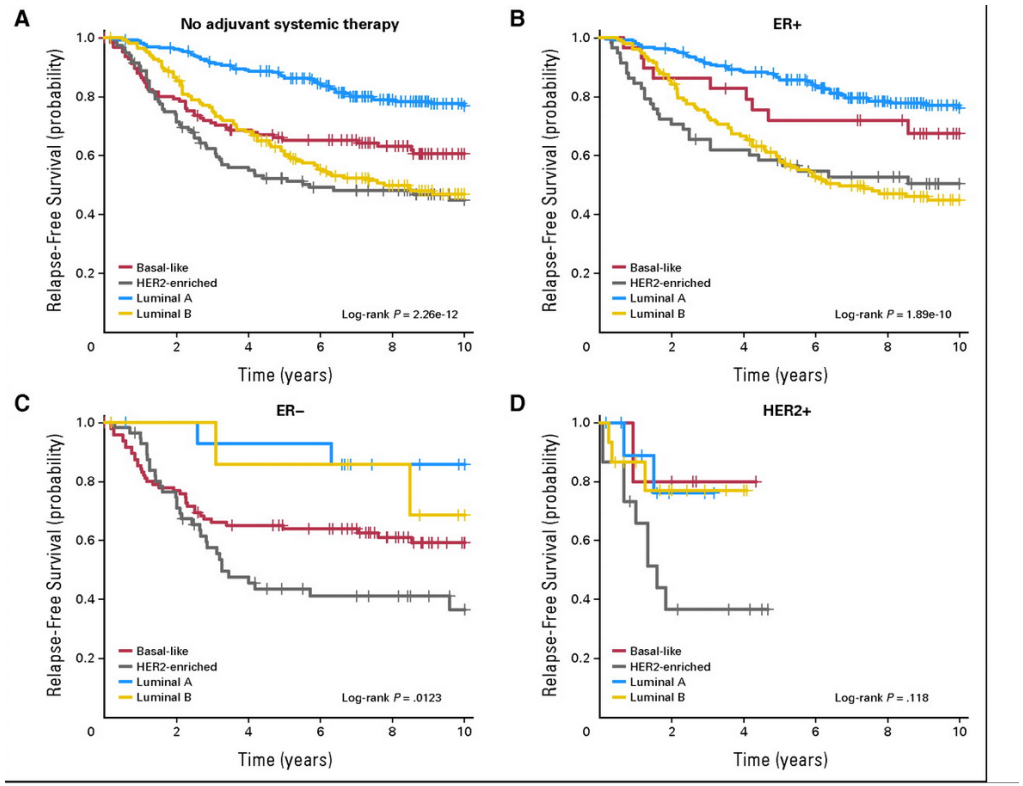
Subtypes of Breast Cancer

- In the past 10 years, global gene expression analyses have identified at least 4 subtypes of breast cancer: Luminal A, Luminal B, Her2-enriched, and basal-like.
- Subgroups differ with respect to risk factors, incidence, baseline prognoses, responses to therapies.
- Want to be able to determine the subtype for a new patient with breast cancer.
- Controversy over the best classifier for this task:
  - PAM50 classifier involves 50 genes.
  - More recent proposal involving three genes.

Subtypes of Breast Cancer

- In the past 10 years, global gene expression analyses have identified at least 4 subtypes of breast cancer: Luminal A, Luminal B, Her2-enriched, and basal-like.
- Subgroups differ with respect to risk factors, incidence, baseline prognoses, responses to therapies.
- Want to be able to determine the subtype for a new patient with breast cancer.
- Controversy over the best classifier for this task:
  - PAM50 classifier involves 50 genes.
  - More recent proposal involving three genes.
- Moving target: nobody knows the "true" subtype!
- Prat et al., Breast Cancer Res Treat, 2012

Why Do We Care About Subtypes?



Citation: Parker et al, Journal of Clinical Oncology, 2009

Proteomics for Ovarian Cancer

Proteomics for Ovarian Cancer

- Ovarian cancer is the leading cause of gynecologic cancer deaths in the USA.

Proteomics for Ovarian Cancer

- Ovarian cancer is the leading cause of gynecologic cancer deaths in the USA.
- Much interest in detecting the cancer at an earlier stage.

Proteomics for Ovarian Cancer

- Ovarian cancer is the leading cause of gynecologic cancer deaths in the USA.
- Much interest in detecting the cancer at an earlier stage.
- In 2002, Petricoin and Liotta – investigators from FDA and NCI – reported in The Lancet that mass spectrometry analysis of circulating serum proteins can be used to discriminate between healthy patients and those with ovarian cancer.

Proteomics for Ovarian Cancer

- Ovarian cancer is the leading cause of gynecologic cancer deaths in the USA.
- Much interest in detecting the cancer at an earlier stage.
- In 2002, Petricoin and Liotta – investigators from FDA and NCI – reported in The Lancet that mass spectrometry analysis of circulating serum proteins can be used to discriminate between healthy patients and those with ovarian cancer.
- Great enthusiasm in the popular press and general public.

Proteomics for Ovarian Cancer

- Ovarian cancer is the leading cause of gynecologic cancer deaths in the USA.
- Much interest in detecting the cancer at an earlier stage.
- In 2002, Petricoin and Liotta – investigators from FDA and NCI – reported in The Lancet that mass spectrometry analysis of circulating serum proteins can be used to discriminate between healthy patients and those with ovarian cancer.
- Great enthusiasm in the popular press and general public.
- Plans were made to begin marketing a test based on the reported diagnostic.

Not So Fast!!

- Independent researchers took a look at the data, which was publicly available, and discovered:
  - inadvertent changes in protocol mid-experiment: i.e. major batch effects.
  - problems with instrument calibration.
  - difference in processing between tumor and normal samples.

Not So Fast!!

- Independent researchers took a look at the data, which was publicly available, and discovered:
  - inadvertent changes in protocol mid-experiment: i.e. major batch effects.
  - problems with instrument calibration.
  - difference in processing between tumor and normal samples.
- In summary: the observed differences between cancer and normal proteomic patterns were attributable to "artifacts of sample processing, not the underlying biology of cancer."

Gene Expression Signatures for Cancer Treatment

Gene Expression Signatures for Cancer Treatment

- In the early 2000's, Joe Nevins, Anil Potti, and other researchers at Duke University began developing expression-based predictors of response to chemotherapy.

Gene Expression Signatures for Cancer Treatment

- In the early 2000's, Joe Nevins, Anil Potti, and other researchers at Duke University began developing expression-based predictors of response to chemotherapy.
- Many (dozens of!) very promising and very high-profile papers were published in Nature Medicine, The Lancet, Journal of Clinical Oncology, and more.

Gene Expression Signatures for Cancer Treatment

- In the early 2000's, Joe Nevins, Anil Potti, and other researchers at Duke University began developing expression-based predictors of response to chemotherapy.
- Many (dozens of!) very promising and very high-profile papers were published in Nature Medicine, The Lancet, Journal of Clinical Oncology, and more.
- Several clinical trials were initiated, using these predictors to direct therapy for cancer patients.

Gene Expression Signatures for Cancer Treatment

- In the early 2000's, Joe Nevins, Anil Potti, and other researchers at Duke University began developing expression-based predictors of response to chemotherapy.
- Many (dozens of!) very promising and very high-profile papers were published in Nature Medicine, The Lancet, Journal of Clinical Oncology, and more.
- Several clinical trials were initiated, using these predictors to direct therapy for cancer patients.
- This research was hailed as a major breakthrough in cancer treatment, and researchers from all over the world tried to use these sorts of techniques in their own labs.

Upon Closer Inspection....

- Using the fact that some of the data were publicly available, independent researchers discovered the following errors (among many others):

Upon Closer Inspection....

- Using the fact that some of the data were publicly available, independent researchers discovered the following errors (among many others):
  - Off-by-one errors in gene lists

Upon Closer Inspection....

- Using the fact that some of the data were publicly available, independent researchers discovered the following errors (among many others):
  - Off-by-one errors in gene lists
  - The same heatmap displayed in multiple (unrelated) papers

Upon Closer Inspection....

- Using the fact that some of the data were publicly available, independent researchers discovered the following errors (among many others):
  - Off-by-one errors in gene lists
  - The same heatmap displayed in multiple (unrelated) papers
  - Genes not measured on the array were reported as being part of the predictor obtained, and as providing evidence for biological plausibility

Upon Closer Inspection....

- Using the fact that some of the data were publicly available, independent researchers discovered the following errors (among many others):
  - Off-by-one errors in gene lists
  - The same heatmap displayed in multiple (unrelated) papers
  - Genes not measured on the array were reported as being part of the predictor obtained, and as providing evidence for biological plausibility
  - Reversal of sensitive/resistant labels

Upon Closer Inspection....

- Using the fact that some of the data were publicly available, independent researchers discovered the following errors (among many others):
  - Off-by-one errors in gene lists
  - The same heatmap displayed in multiple (unrelated) papers
  - Genes not measured on the array were reported as being part of the predictor obtained, and as providing evidence for biological plausibility
  - Reversal of sensitive/resistant labels
- A shocking paper published by Baggerly and Coombes in Annals of Applied Statistics, detailing all of the errors made: "One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common."

What Went Wrong?

A blasé approach to high-dimensional data analysis:

What Went Wrong?

A blasé approach to high-dimensional data analysis:

- Need to have a proper independent test set, that you simply cannot peek at under any circumstances!

What Went Wrong?

A blasé approach to high-dimensional data analysis:

- Need to have a proper independent test set, that you simply cannot peek at under any circumstances!
- Need to have clearly documented code that contains all steps of the analysis, from start to finish. You must be able to share this code with independent researchers, and you must be confident that your code is correct. If not, then your work isn't ready for prime time.

The Stakes are High!

At Duke:

- Dozens of papers retracted;
- Careers and reputations ruined;
- Patients endangered through unethical clinical trials.

Plus, a 60 Minutes special feature and an Institute of Medicine Committee!!!