Statistical Challenges in Proteomics

Making Sense of Two-Dimensional Electrophoretic Data

Françoise Seillier-Moiseiwitsch

Anindya Roy

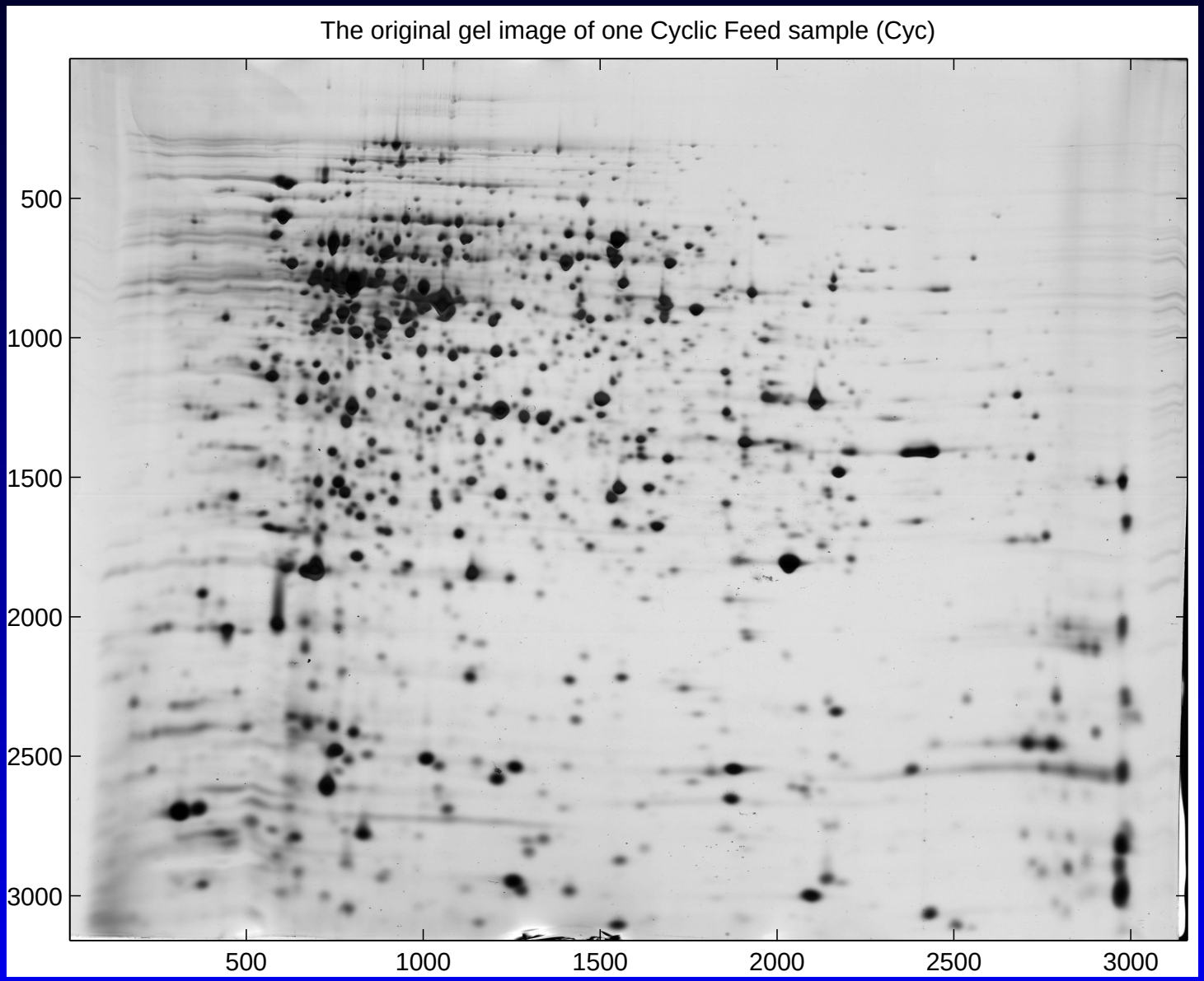
Yaming Hang

Florian Potra

Xing Liu

Department of Mathematics & Statistics

University of Maryland, Baltimore County


The original gel image of one Cyclic Feed sample (Cyc)
500
1000
1500
2000
2500
3000
500
1000
1500
2000
2500
3000

Outline

- Description of the Technology
  - 2-D polyacrylamide gel electrophoresis
    - separate thousands of proteins
    - (> 2,000 for mammalian cell sample)

Outline

- Description of the Technology
  - 2-D polyacrylamide gel electrophoresis
    - separate thousands of proteins
    - (> 2,000 for mammalian cell sample)
- State-of-the-Art Analytical Methods
  - noise reduction
  - spot identification
  - feature selection

Outline

- Description of the Technology
  - 2-D polyacrylamide gel electrophoresis
    - separate thousands of proteins
    - ($>$ 2,000 for mammalian cell sample)
- State-of-the-Art Analytical Methods
  - noise reduction
  - spot identification
  - feature selection
- Statistical Approaches
  - spike removal
  - streak removal
  - gel alignment
  - modeling

The Technology

First dimension isoelectric focusing

- proteins focused electrophoretically in pH gradient
- stop moving when at position with no net charge
  isoelectric point

Second dimension molecular mass

- proteins coated with sodium dodecylsulphate (SDS)
  $\longrightarrow$ same charge density
- separated orthogonally by electrophoresis on
  polyacrylamide gel

independent dimensions

The Technology (ctd)

separated proteins stained with fluorescent dyes

image of displayed proteins = proteome

digital scanning into database

Some Challenges

- gel reproducibility
  same quality from day to day and from lab to lab
- losses due to hydrophobic interactions between some proteins and gel
- removal of nucleic acids
  streaks, artifactual migration
- low abundance protein
  copy number for detection on gel?

Analytical Methods

implemented in MELANIE package

- Spot detection

  non-parametric method based on Laplacian and second derivatives

  $I(x, y)$ = 2-D image

  $\mathbf{p} = (x, y)$ = point on image

  S_i = spot

  T = saturation threshold

  $\Delta I(\mathbf{p})$ = Laplacian

$$= -\left(\frac{\partial^2}{\partial x^2} I(\mathbf{p}) + \frac{\partial^2}{\partial y^2} I(\mathbf{p})\right)$$

Analytical Methods (ctd)

Is **p** part of a spot?
l, r, c small positive thresholds

- $I(\mathbf{p}) < T$

$$\mathbf{p} \in S_i \iff \min\left(\frac{\partial^2}{\partial x^2} I(\mathbf{p}) - r, \frac{\partial^2}{\partial y^2} I(\mathbf{p}) - c\right) > 0$$

when $-\Delta I(\mathbf{p}) - l \geq 0$

- $I(\mathbf{p}) > T$

$$\mathbf{p} \in S_i \iff \min\left(\frac{\partial^2}{\partial x^2} I(\mathbf{p}), \frac{\partial^2}{\partial y^2} I(\mathbf{p})\right) > 0$$

Analytical Methods (ctd)

- Spot quantification

Direct method

spot area = number of pixels $\times$ pixel area

spot optical density = $\max_{x,y \in \text{spot}} I(x,y)$

spot volume = $\sum_{x,y \in \text{spot}} I(x,y)$

Gaussian curve fitting

Analytical Methods (ctd)

- Image alignment

  polynomial image warping

$$(x, y) \rightarrow (u(x), v(y))$$

  (x, y) = pixel coordinates in original image
  (u, v) = pixel coordinates in warped image
  = 1st-, 2nd- or 3rd-order polynomials estimated via least-squares criterion
  – select landmarks on each image
  – choose one gel as reference
  – estimate parameters by summing over landmarks

Analytical Methods (ctd)

- Spot matching
  - for each spot, select cluster of neighbor spots
    central spot = primary spot
    surrounding spots = secondary spots
    spot $\in$ cluster if centroid in circle of fixed radius
    radius depends on image dimension, number of spots, minimum number of spots in cluster
  - match highest-intensity clusters
    primary spots
  - compare clusters
    probabilistic similarity measure:
    probability of next random hit within a cluster where $m-1$ spots have been matched

Analytical Methods (ctd)

- Spot matching (ctd)

  – consistency check for possible mismatching

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} t_x \\ t_y \end{pmatrix} + \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

  check $L = AD - BC \approx 1$ rotation
  for each primary cluster parameters estimated from 3 matched spots

  $L = 1.0 \pm 0.25$ and $\theta \pm 10$ degrees

  $\rightarrow$ *good match*

  – transformation to match remaining spots
  from good matches in 2 clusters estimate
  $A, \cdots, D$ by least-square method

Analytical Methods (ctd)

- Creating synthetic gels
  merge gel images to get master gel
  – select reference gel
  spot positions in reference gel
  = spot position in synthetic gel
  – check spots on reference gel well matched to spots on 2 other gels
  → triangles of matched spots (= starting groups)
  – extend starting groups by adding spots
  connectivity test:
  spot matched with at least one other spot in initial group

Analytical Methods (ctd)

- Creating synthetic gels (ctd)

  – when all spots on reference gel considered, create additional groups with spots on 2nd gel which are not part of group and repeat with other gels

  – synthetic gel contains same number of spots as determined groups

  representative spots

  position: from reference gel if group has spot on reference gel
  closest spot + translation otherwise

  intensity: average for spots in group

  shape: shape of spot in group closest in area to average of group

Analytical Methods (ctd)

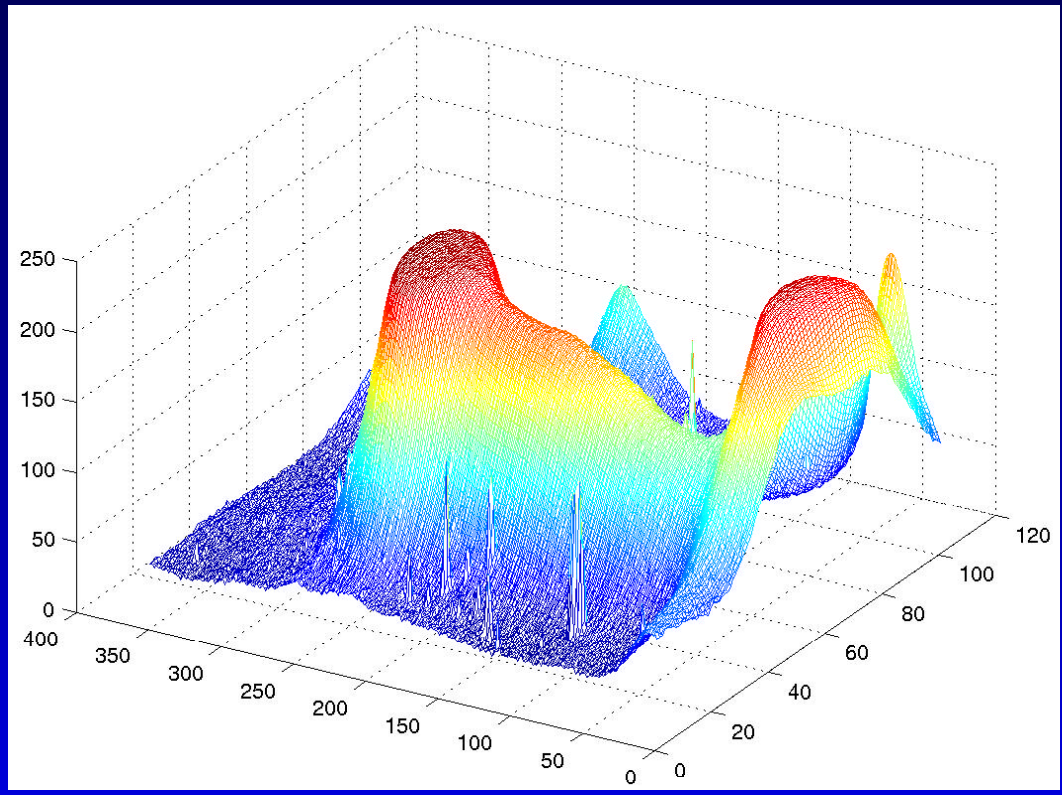
- Filtering gel images
  - Gaussian smoothing
  - diffusion smoothing
  - polynomial smoothing
  - daptive smoothing
    preserves significant discontinuities

Analytical Methods (ctd)

- Background filtering
  - global minimum pixel value
  - estimate background outside spots with 3rd-order polynomial

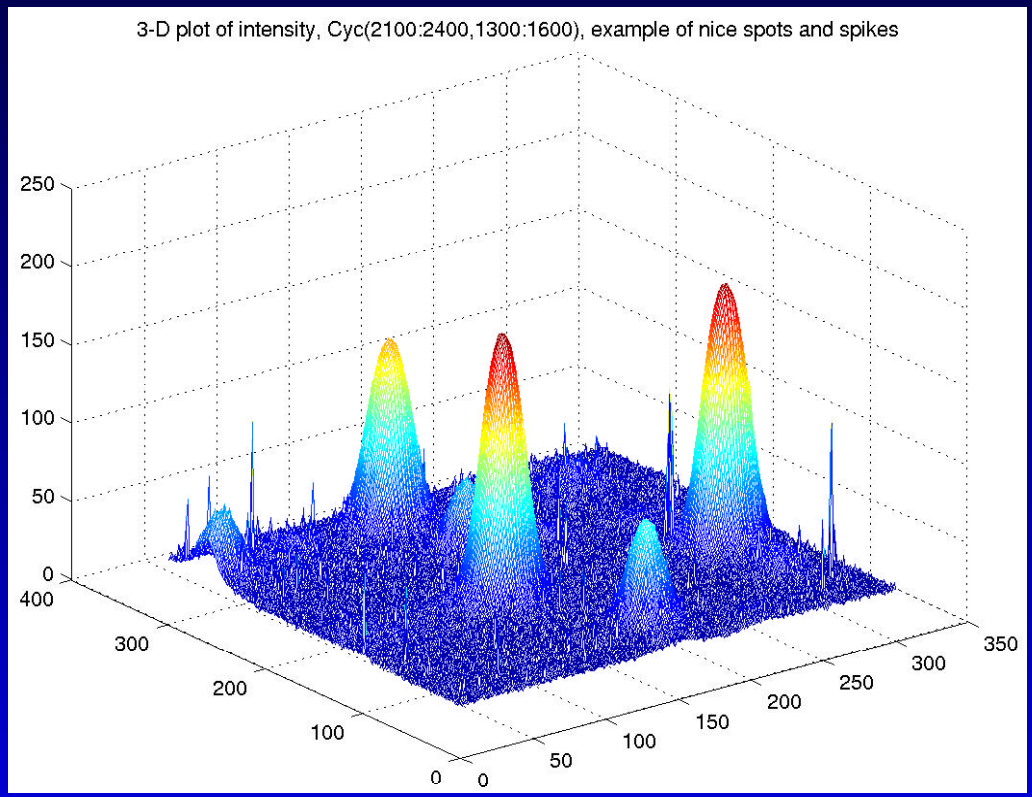
Statistical Approaches

Gaussian?



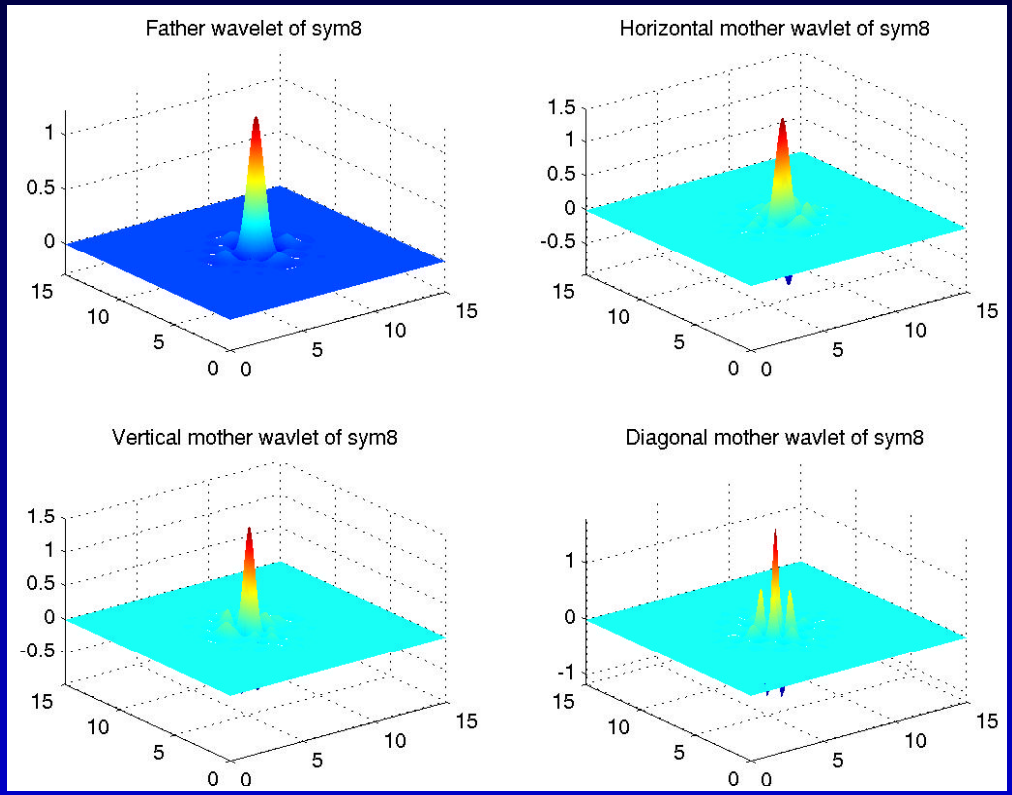
Statistical Approaches (ctd)

- Spike Removal



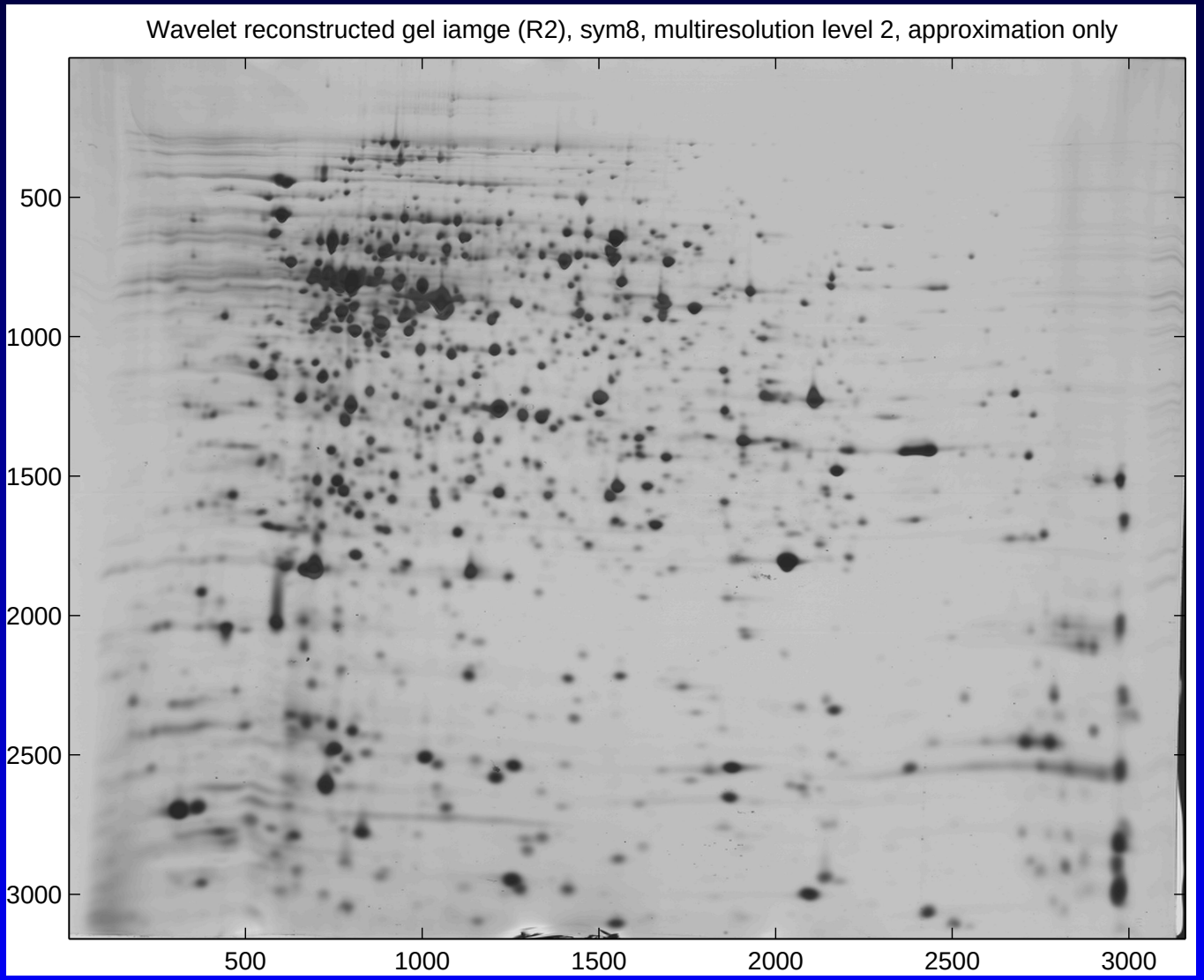
Statistical Approaches (ctd)

Wavelet methods



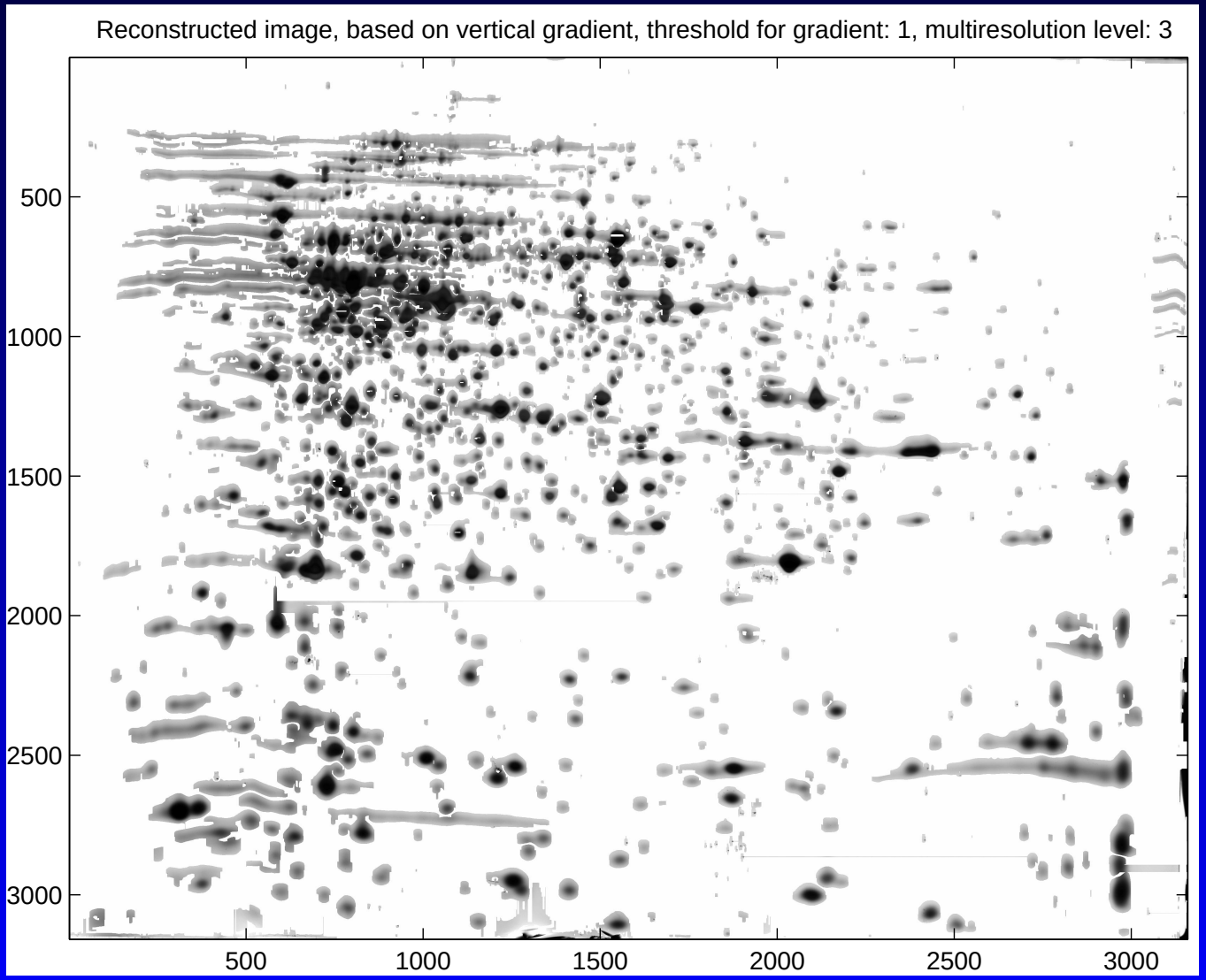
Statistical Approaches (ctd)

- Streak Removal



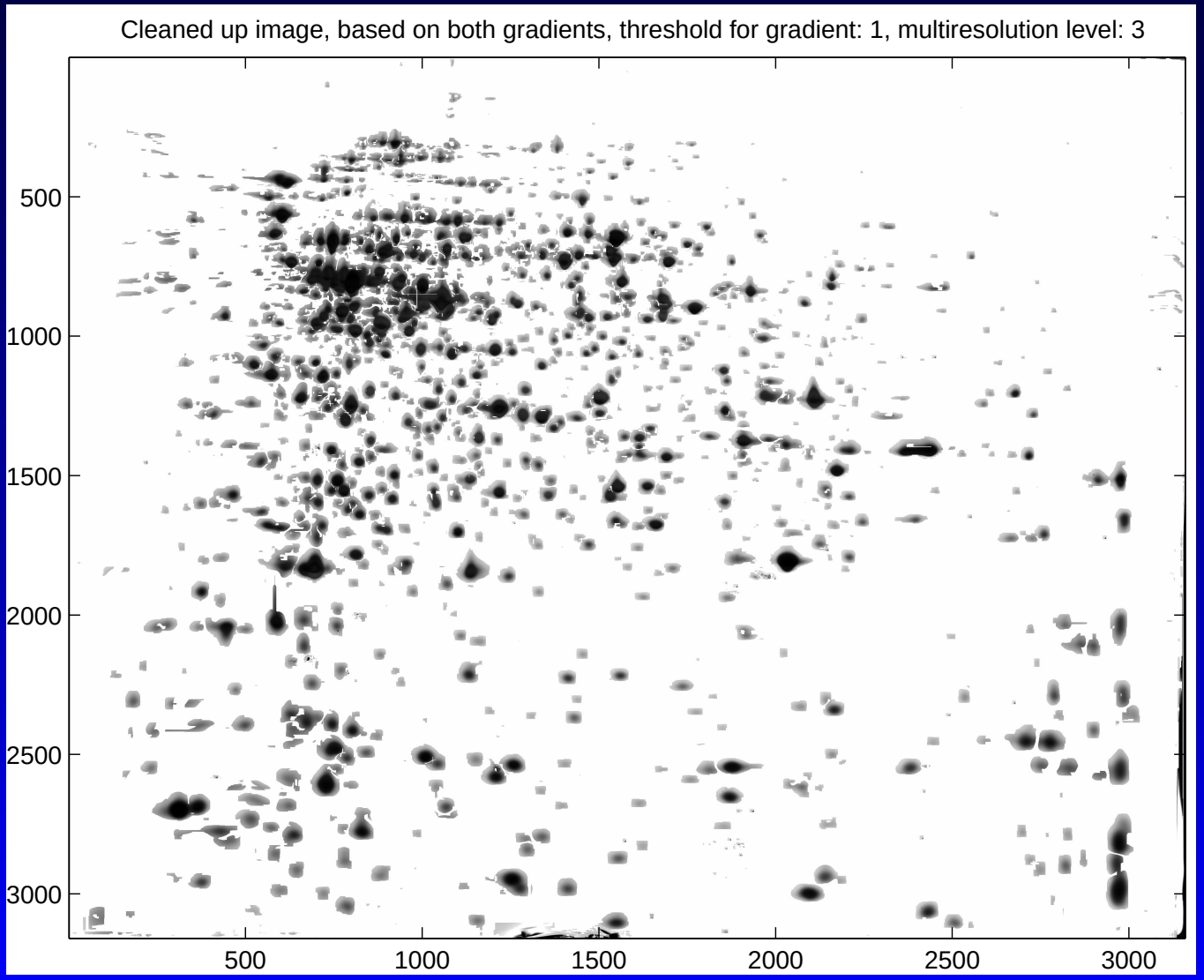
Statistical Approaches (ctd)

- Streak Removal (ctd)



Statistical Approaches (ctd)

- Streak Removal (ctd)



Statistical Approaches (ctd)

- Gel Alignment

N gels with M points on each gel
Find transformations
$T_i : R^2 \rightarrow R^2$, for gel i ($i = 1, 2, \ldots, N$)
describing changes for all spots on the gel.
Linear transformations for every gel:
$T_i(x) = A_i x + b_i$, where

$$A_i = \begin{pmatrix} \alpha_i & \beta_i \\ \gamma_i & \delta_i \end{pmatrix}, b_i = \begin{pmatrix} \varphi_i \\ \psi_i \end{pmatrix}, i = 1, 2, \ldots, N$$

Statistical Approaches (ctd)

- Gel Alignment (ctd)

  Position of center of spot j on gel i :
  $l_{ij}, i = 1, 2, \ldots, N, j = 1, 2, \ldots, M$, i.e.

  $$l_{ij} = \begin{pmatrix} l_{ij}^1 \\ l_{ij}^2 \end{pmatrix}$$

  Transformed l_{ij}

  $$\begin{pmatrix} \theta_{ij}^1 \\ \theta_{ij}^2 \end{pmatrix} = \theta_{ij} = T_i(l_{ij})$$

  $$\theta_{ij}^1 = \alpha_i l_{ij}^1 + \beta_i l_{ij}^2 + \varphi_i$$

  $$\theta_{ij}^2 = \gamma_i l_{ij}^1 + \delta_i l_{ij}^2 + \psi_i$$

  True location of spot j : l_j

Statistical Approaches (ctd)

- Gel Alignment (ctd)

  Restriction on transformed points:
  – not be too far away from the true location
  $\|\theta_{ij} - l_i\|_\infty \leq \epsilon_{ij}$, i.e.

  $$-\epsilon_{ij} \leq \theta^1_{ij} - l^1_i \leq \epsilon_{ij}, i = 1, 2, \ldots, N, j = 1, 2, \ldots, M$$

  $$-\epsilon_{ij} \leq \theta^2_{ij} - l^2_i \leq \epsilon_{ij}, i = 1, 2, \ldots, N, j = 1, 2, \ldots, M$$

  where ϵ_{ij} is a given error bound.
  – transformation as close as possible to identity transformation

Statistical Approaches (ctd)

- Gel Alignment (ctd)

  Objective function:

$$\sum_{i=1}^{N}(\|A_i - I\|^2 + \|b_i\|^2)$$

$$+ \text{ penalty term}$$

  weighted sum distance between true location and average spot centers

  Constraint: weighted sum of distances between true location and transform points

Statistical Approaches (ctd)

- Gel Alignment (ctd)

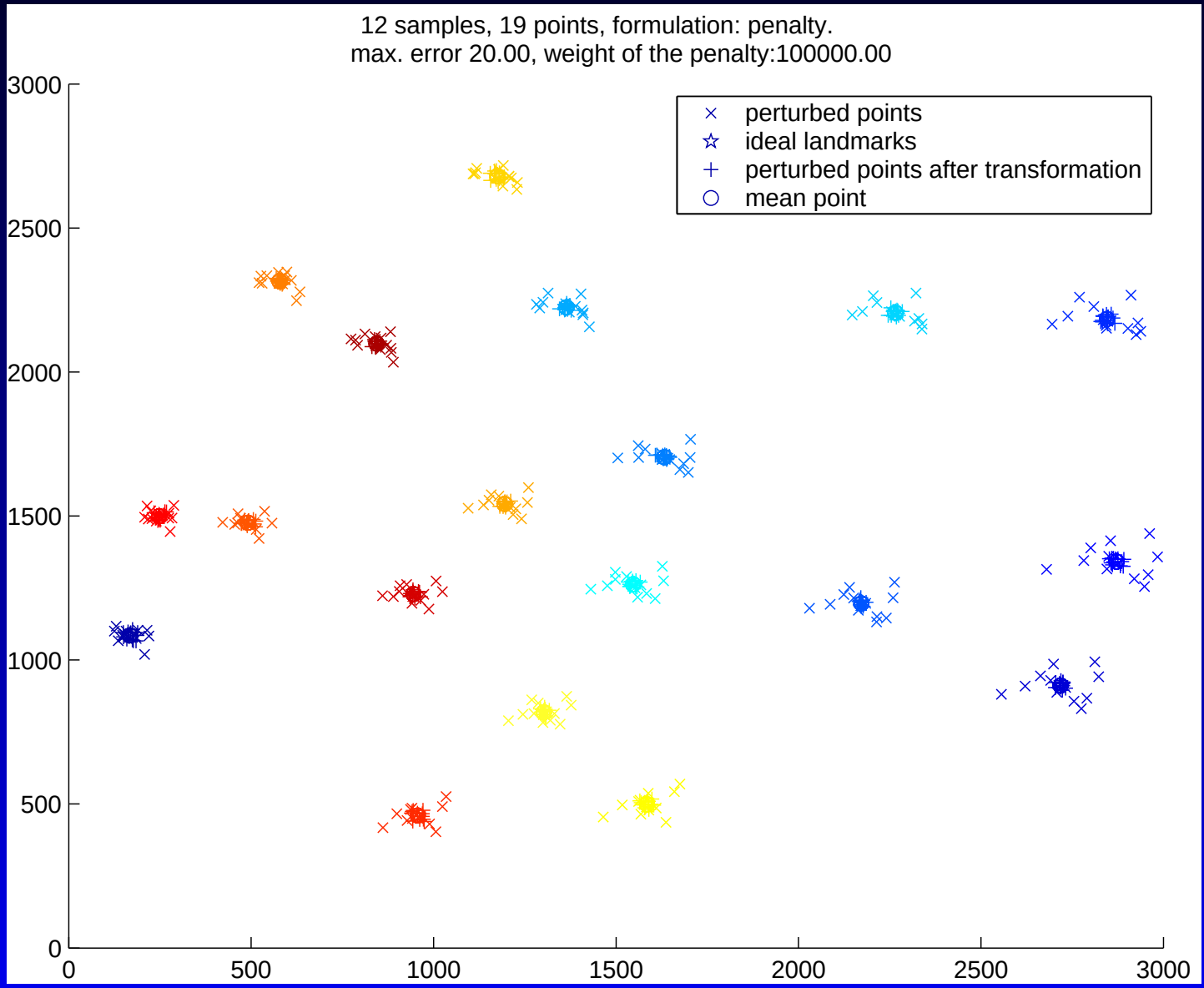
  Quadratic Programming

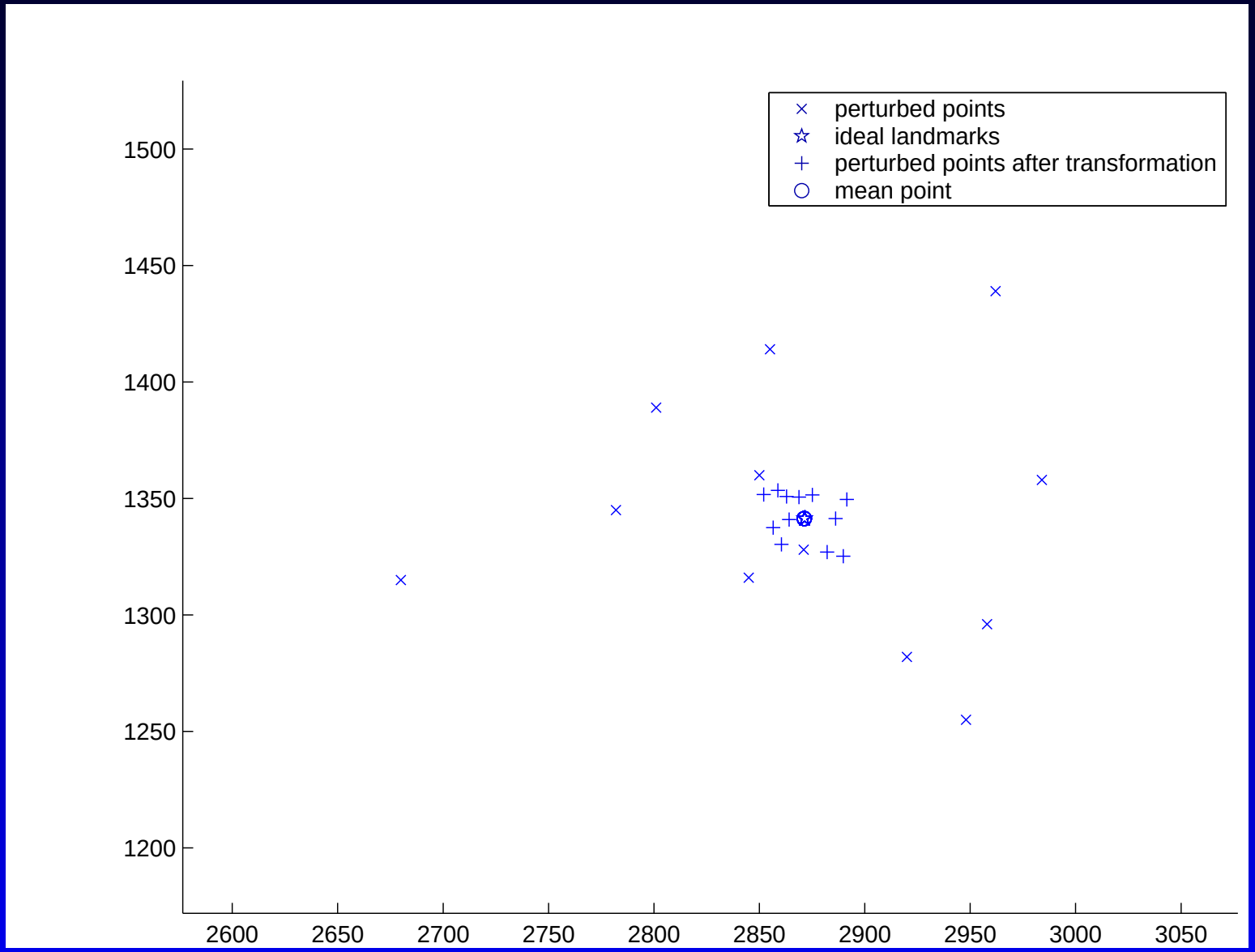
  Given data l_{ij} ($i = 1, 2, \ldots, N, j = 1, 2, \ldots, M$), maximum allowable error ϵ_{ij}, and a weight w, first compute the mean of each spot center on different images ml_j, then solve

$$\min_{\alpha_i,\beta_i,\gamma_i,\delta_i,\varphi_i,\psi_i,l_j} \sum_{i=1}^{N} \left[(\alpha_i - 1)^2 + \beta_i^2 + \gamma_i^2 + (\delta_i - 1)^2 \right.$$

$$\left. + \varphi_i^2 + \psi_i^2 \right] + w \sum_{j=1}^{M} \left[(l_j^1 - m l_j^1)^2 + (l_j^2 - m l_j^2)^2\right]$$

$$\text{s.t.} \quad -\epsilon_{ij} \leq \alpha_i l_{ij}^1 + \beta_i l_{ij}^2 + \varphi_i - l_j^1 \leq \epsilon_{ij}$$
$$-\epsilon_{ij} \leq \gamma_i l_{ij}^1 + \delta_i l_{ij}^2 + \psi_i - l_j^2 \leq \epsilon_{ij}$$


12 samples, 19 points, formulation: penalty.
max. error 20.00, weight of the penalty:100000.00
perturbed points
ideal landmarks
perturbed points after transformation
mean point
0
500
1000
1500
2000
2500
3000


perturbed points
ideal landmarks
perturbed points after transformation
mean point
1500
1450
1400
1350
1300
1250
1200
2600
2650
2700
2750
2800
2850
2900
2950
3000
3050

Statistical Approaches (ctd)

- Statistical Modeling

  – Analysis of variance for spot volumes

  Mixed-effect model:

  fixed group effect
  random individual effect
  spatial correlation structure

  – Feature selection based on hypothesis testing of groups of wavelet coefficients