Empirical Bayes Unfolding of Elementary Particle Spectra at the Large Hadron Collider

Mikael Kuusela

Institute of Mathematics,
EPFL

Statistics Seminar,
University of Bristol

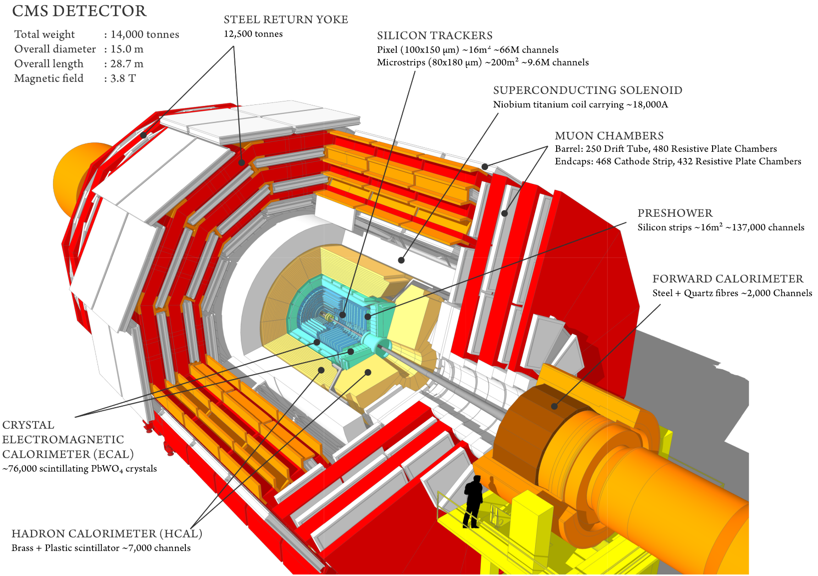
June 13, 2014

Joint work with Victor M. Panaretos

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

CERN and the Large Hadron Collider

CMS Experiment at the LHC



Statistics at CERN

- Hypothesis testing / interval estimation with a large number of nuisance parameters
  - Higgs boson, supersymmetry, beyond Standard Model physics,...
- Nonparametric multiple regression
  - Energy response calibration
- Statistical inverse problems
  - Unfolding
- Classification
  - Improve S/B ratio, particle identification, triggering
- Pattern recognition
  - Particle tracking

The Unfolding Problem

- Any measurement carried out at the LHC is affected by the finite resolution of the particle detectors
- This causes the observed spectrum of events to be "smeared" or "blurred" with respect to the true one
- The *unfolding problem* is to estimate the true spectrum using the smeared observations
  - Mathematically closely related to deblurring in optics and tomographic image reconstruction in medical imaging

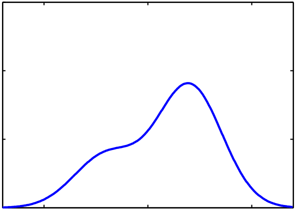
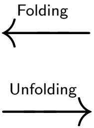


Figure : Smeared spectrum



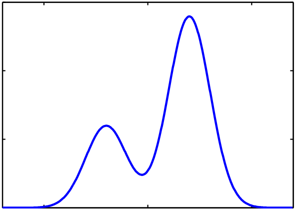


Figure : True spectrum

Unfolding is an Ill-Posed Inverse Problem

- The main issue in unfolding is the ill-posedness of the mapping from the true spectrum to the smeared spectrum
  - The (pseudo)inverse of this mapping is very sensitive to small perturbations of the data
- Need to regularize the problem by introducing additional information about plausible solutions
- Current "state-of-the-art":
  1. EM iteration with early stopping
  2. Generalized Tikhonov regularization
- Two major challenges:
  1. How to choose the regularization strength?
  2. How to quantify the uncertainty of the solution?
- In this talk, we propose an empirical Bayes unfolding framework for tackling these issues

Problem Formulation Using Poisson Point Processes (1)

- The appropriate mathematical model for unfolding is that of *indirectly observed Poisson point processes*
- A random measure M is a *Poisson point process* with intensity function f and state space E iff
  1. $M(B) \sim \mathrm{Poisson}(\lambda(B))$, where $\lambda(B) = \int_B f(s)\,\mathrm{d}s$, for every Borel set $B \subset E$
  2. $M(B_1), \ldots, M(B_n)$ are independent random variables for disjoint Borel sets $B_1, \ldots, B_n \subset E$
- The intensity function f uniquely characterizes the law of M
  - I.e., all the information about the behavior of M is contained in f

Problem Formulation Using Poisson Point Processes (2)

- Let M and N be two Poisson point processes with intensities f and g and state spaces E and F, respectively
- Assume that M represents the true, particle-level events and N the smeared, detector-level events
- Then

$$g(t) = (Kf)(t) = \int_E k(t,s) f(s) \, \mathrm{d}s,$$

where the smearing kernel k represents the response of the detector and is given by

$$k(t,s) = p(Y_i = t | X_i = s, i\text{th event observed}) P(i\text{th event observed} | X_i = s),$$

where X_i is the ith true event and Y_i the corresponding smeared event
- Task: Estimate f given a single realization of the process N

Empirical Bayes Unfolding

- We propose to estimate f based on the following key principles:
  1. Discretization of the true intensity f using a **cubic B-spline basis expansion**, that is,
  $$f(s) = \sum_{j=1}^{p} \beta_j B_j(s),$$
  where B_j, $j = 1, \ldots, p$, are the B-spline basis functions
  2. **Posterior mean estimation** of the B-spline coefficients $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^{\mathrm{T}}$
  3. **Empirical Bayes selection** of the scale δ of the regularizing smoothness prior $p(\boldsymbol{\beta}|\delta)$
  4. Frequentist uncertainty quantification and bias correction using the **parametric bootstrap**

Discretization of the Problem

- Let $\{F_i\}_{i=1}^n$ be a partition of the smeared space F with n intervals
- Let $y_i = N(F_i)$ be the number of points observed in interval F_i
  - I.e., we record the observations into a histogram $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$
- Then

$$\begin{aligned} \mathrm{E}(y_i|\boldsymbol{\beta}) &= \int_{F_i} g(t)\,\mathrm{d}t = \int_{F_i}\int_E k(t,s) f(s)\,\mathrm{d}s\,\mathrm{d}t \\ &= \sum_{j=1}^{p} \Bigg(\underbrace{\int_{F_i}\int_E k(t,s) B_j(s)\,\mathrm{d}s\,\mathrm{d}t}_{:=K_{i,j}} \Bigg) \beta_j = \sum_{j=1}^{p} K_{i,j}\beta_j \end{aligned}$$

- Hence, we need to solve the Poisson regression problem

$$\mathbf{y}|\boldsymbol{\beta} \sim \mathrm{Poisson}(\mathbf{K}\boldsymbol{\beta})$$

for an ill-conditioned matrix $\mathbf{K}$

Bayesian Estimation of the Spline Coefficients

- Posterior for $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}|\mathbf{y},\delta) = \frac{p(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\delta)}{p(\mathbf{y}|\delta)}, \quad \boldsymbol{\beta} \in \mathbb{R}_+^p,$$

  where the likelihood is given by

$$p(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\left(\sum_{j=1}^{p} K_{i,j}\beta_j\right)^{y_i}}{y_i!} e^{-\sum_{j=1}^{p} K_{i,j}\beta_j}, \quad \boldsymbol{\beta} \in \mathbb{R}_+^p$$

- We regularize the problem using the Gaussian smoothness prior

$$p(\boldsymbol{\beta}|\delta) \propto \exp\left(-\delta\|f''\|_2^2\right) = \exp\left(-\delta\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Omega}\boldsymbol{\beta}\right), \quad \boldsymbol{\beta} \in \mathbb{R}_+^p,$$

  with $\delta > 0$ and $\Omega_{i,j} = \int_E B_i''(s)B_j''(s)\,\mathrm{d}s$
  - This becomes a proper pdf once we impose Aristotelian boundary conditions
- We use a single-component Metropolis–Hastings algorithm to sample from the posterior
  - The univariate proposal densities are chosen to approximate the full conditionals $p(\beta_k|\boldsymbol{\beta}_{-k},\mathbf{y},\delta)$ of the Gibbs sampler as proposed by Saquib et al. (1998)

Empirical Bayes Estimation of the Hyperparameter

- We propose choosing the hyperparameter δ (i.e. the regularization parameter) via marginal maximum likelihood:

$$\hat{\delta} = \hat{\delta}(\mathbf{y}) = \underset{\delta>0}{\arg\max}\ p(\mathbf{y}|\delta) = \underset{\delta>0}{\arg\max} \int_{\mathbb{R}_+^p} p(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\delta)\,\mathrm{d}\boldsymbol{\beta}$$

- The marginal maximum likelihood estimate $\hat{\delta}$ is found using a Monte Carlo expectation-maximization algorithm (Geman and McClure, 1985, 1987; Saquib et al., 1998):

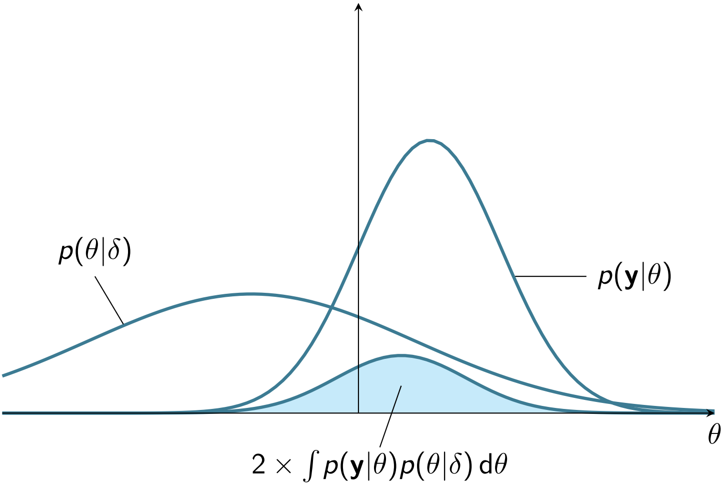
  E-step: Sample $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(S)}$ from the posterior $p(\boldsymbol{\beta}|\mathbf{y}, \delta^{(t)})$ and compute $Q(\delta; \delta^{(t)}) = \frac{1}{S}\sum_{s=1}^{S} \log p(\boldsymbol{\beta}^{(s)}|\delta)$

  M-step: Set $\delta^{(t+1)} = \arg\max_{\delta>0} Q(\delta; \delta^{(t)})$

- The spline coefficients $\boldsymbol{\beta}$ are then estimated using the mean of the empirical Bayes posterior: $\hat{\boldsymbol{\beta}} = \mathrm{E}(\boldsymbol{\beta}|\mathbf{y}, \hat{\delta})$
- The estimated intensity is $\hat{f}(s) = \sum_{j=1}^{p} \hat{\beta}_j B_j(s)$

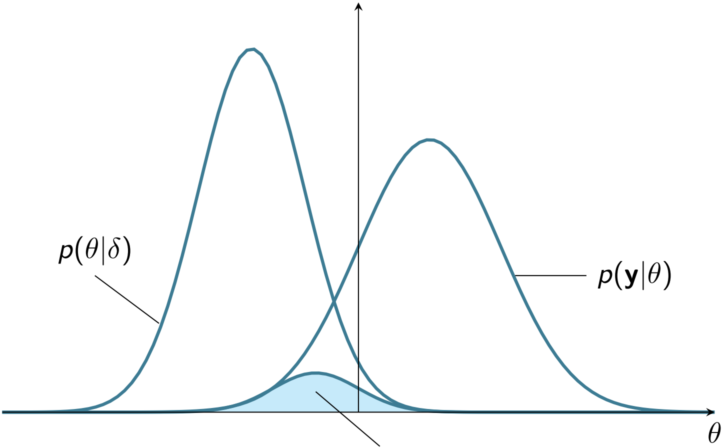
What Does Empirical Bayes Do?

$$\hat{\delta} = \arg\max_{\delta>0}\, p(\mathbf{y}|\delta) = \arg\max_{\delta>0} \int p(\mathbf{y}|\theta)p(\theta|\delta)\,\mathrm{d}\theta$$



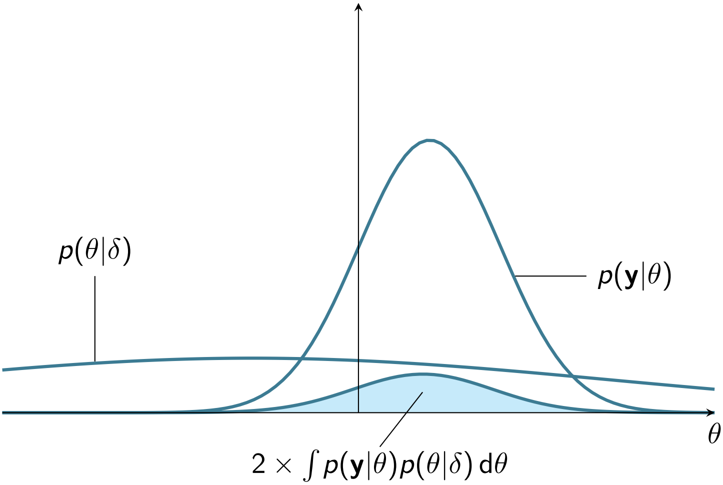
What Does Empirical Bayes Do?

$$\hat{\delta} = \underset{\delta>0}{\arg\max}\ p(\mathbf{y}|\delta) = \underset{\delta>0}{\arg\max} \int p(\mathbf{y}|\theta)p(\theta|\delta)\,\mathrm{d}\theta$$



What Does Empirical Bayes Do?

$$\hat{\delta} = \arg\max_{\delta>0} \, p(\mathbf{y}|\delta) = \arg\max_{\delta>0} \int p(\mathbf{y}|\theta)p(\theta|\delta)\,\mathrm{d}\theta$$



Empirical Bayes vs. Hierarchical Bayes

- Hierarchical Bayes is a natural alternative for empirical Bayes
- But need to choose the hyperprior $p(\delta)$
  - It is a priori unclear how this should be done
  - Different choices can result in non-negligible differences in the posterior
  - The choice is not necessarily invariant under reparametrizations
- Empirical Bayes on the other hand:
  - Chooses a unique, "best" regularizer among the family of priors $\{p(\boldsymbol{\beta}|\delta)\}_{\delta>0}$
  - Requires *only* the choice of the family $\{p(\boldsymbol{\beta}|\delta)\}_{\delta>0}$
  - Is by construction transformation invariant
- Empirical Bayes has become part of the standard methodology in generalized additive models (Wood, 2011) and Gaussian processes (Rasmussen and Williams, 2006)
  - What about inverse problems?

Uncertainty Quantification and Bias Correction (1)

- The credible intervals of the empirical Bayes posterior $p(\boldsymbol{\beta}|\mathbf{y}, \hat{\delta})$ could in principle be used to make confidence statements about f
  - But due to the data-driven choice of the prior, these intervals lose their subjective Bayesian interpretation
  - Furthermore, their frequentist properties are poorly understood
- Instead, we propose using the parametric bootstrap to construct frequentist confidence bands for f:
  1. Obtain a resampled observation $\mathbf{y}^*$
  2. Rerun the MCEM algorithm with $\mathbf{y}^*$ to find $\hat{\delta}^* = \hat{\delta}(\mathbf{y}^*)$
  3. Compute $\hat{\boldsymbol{\beta}}^* = \mathrm{E}(\boldsymbol{\beta}|\mathbf{y}^*, \hat{\delta}^*)$
  4. Obtain $\hat{f}^*(s) = \sum_{j=1}^{p} \hat{\beta}_j^* B_j(s)$
  5. Repeat R times
- The bootstrap sample $\{\hat{f}^{*(r)}\}_{r=1}^{R}$ is then used to compute approximate frequentist confidence intervals for $f(s)$ for each $s \in E$
- This procedure also takes into account uncertainty regarding the choice of the hyperparameter δ

Uncertainty Quantification and Bias Correction (2)

- One can envisage various ways of obtaining the resampled observations $\mathbf{y}^*$ and of using the bootstrap sample $\{\hat{f}^{*(r)}\}_{r=1}^{R}$ to compute approximate frequentist confidence bands
- We propose using:

  Resampling: $\mathbf{y}^* \overset{\text{i.i.d.}}{\sim} \mathrm{Poisson}(\mathbf{K}\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}} = \mathrm{E}(\boldsymbol{\beta}|\mathbf{y}, \hat{\delta})$

  Intervals: Pointwise $1 - 2\alpha$ basic bootstrap intervals, given by

$$[2\hat{f}(s) - \hat{f}^*_{1-\alpha}(s), 2\hat{f}(s) - \hat{f}^*_{\alpha}(s)]$$

- Here $\hat{f}^*_{\alpha}(s)$ denotes the α-quantile of the bootstrap sample evaluated at point $s \in E$
- The bootstrap may also be used to correct for the unavoidable bias in the point estimate $\hat{f}$
- Bootstrap estimate of the bias: $\widehat{\mathrm{bias}}^*(\hat{f}(s)) = \frac{1}{R}\sum_{r=1}^{R} \hat{f}^{*(r)}(s) - \hat{f}(s)$
- Bias-corrected point estimate: $\hat{f}_{\mathrm{BC}}(s) = \hat{f}(s) - \widehat{\mathrm{bias}}^*(\hat{f}(s))$

Demonstration: Setup

- True intensity

$$f(s) = \lambda_{\text{tot}} \left\{ \pi_1 \mathcal{N}(s|-2,1) + \pi_2 \mathcal{N}(s|2,1) + \pi_3 \frac{1}{|E|} \right\},$$

with $\pi_1 = 0.2, \pi_2 = 0.5$ and $\pi_2 = 0.3$

- Smeared intensity

$$g(t) = \int_E \mathcal{N}(t-s|0,1) f(s) \, \mathrm{d}s$$

- $E = F = [-7, 7]$, discretized using $n = 40$ histogram bins and $p = 30$ B-spline basis functions
- The condition number of the smearing matrix $\mathbf{K}$ is $2.6 \cdot 10^8$
  - $\Rightarrow$ Problem severely ill-posed!

Demonstration: Empirical Bayes Unfolding, $\lambda_{\text{tot}} = 20\,000$

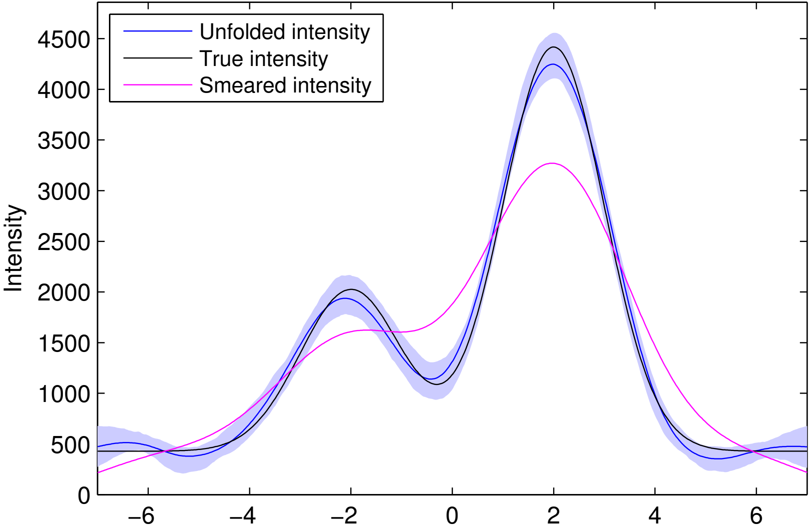


Figure : Empirical Bayes unfolding, $\lambda_{\text{tot}} = 20\,000$, 95 % pointwise basic intervals

Demonstration: Empirical Bayes Unfolding, $\lambda_{\text{tot}} = 1\,000$

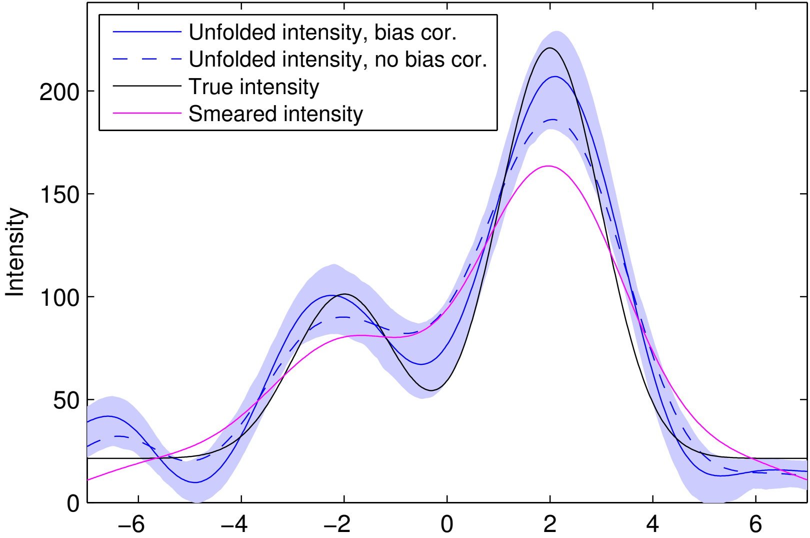


Figure : Empirical Bayes unfolding, $\lambda_{\text{tot}} = 1\,000$, 95 % pointwise basic intervals

Demonstration: Hierarchical Bayes Unfolding, $\lambda_{\text{tot}} = 1\,000$

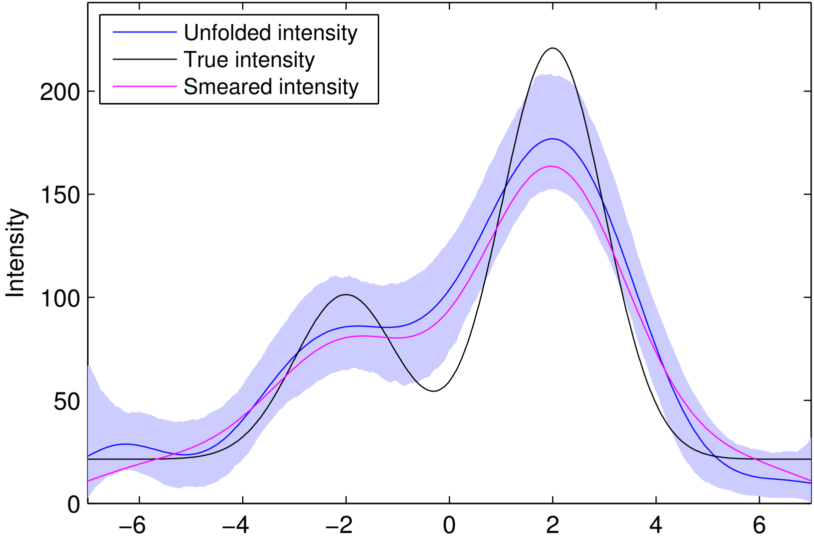


Figure : Hierarchical Bayes, $\delta \sim \mathrm{Gamma}(1, 0.05)$, 95 % credible intervals

Demonstration: Hierarchical Bayes Unfolding, $\lambda_{\mathrm{tot}} = 1\,000$

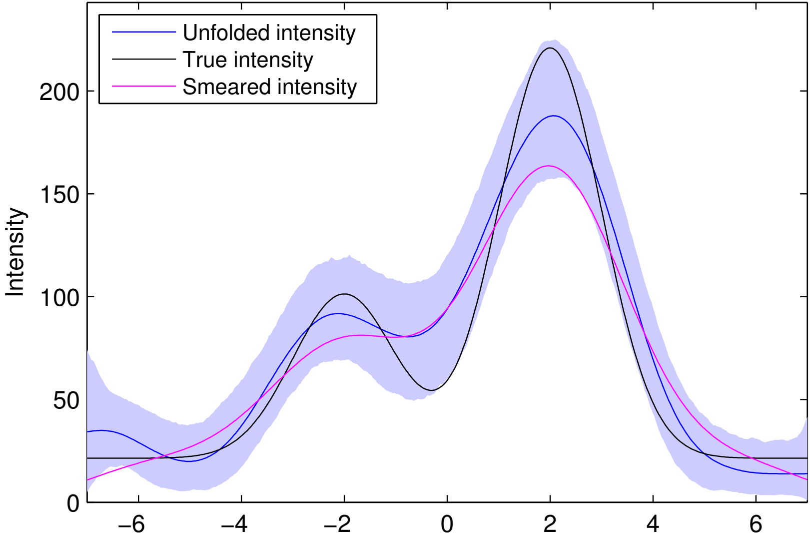


Figure : Hierarchical Bayes, $\delta \sim \mathrm{Gamma}(0.001, 0.001)$, 95 % credible intervals

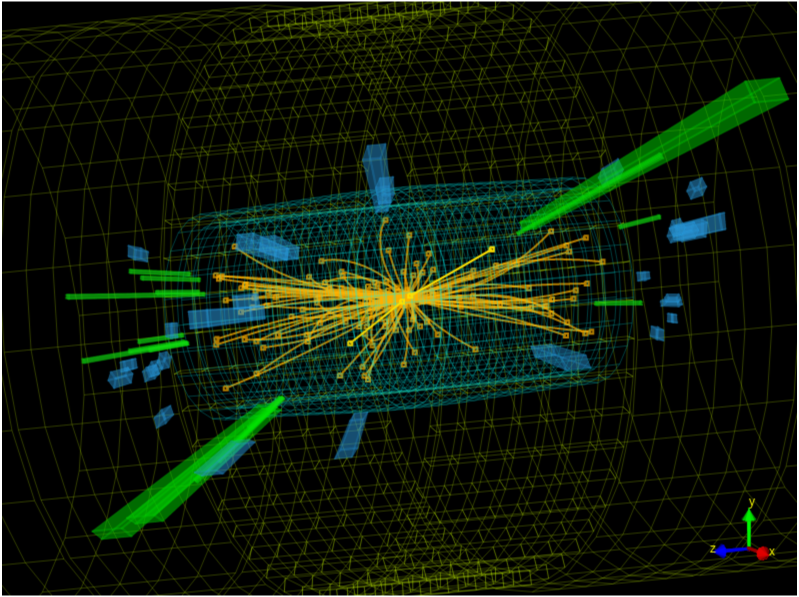
$Z \to e^+e^-$: Setup

- We demonstrate empirical Bayes unfolding with real data by unfolding the $Z \to e^+e^-$ invariant mass spectrum measured in CMS
- The data are published in Chatrchyan et al. (2013) and correspond to integrated luminosity of 4.98 fb^{-1} collected in 2011 at $\sqrt{s} = 7\ \mathrm{TeV}$
- 67 778 "high quality" electron-positron pairs with invariant masses 65–115 GeV in 0.5 GeV bins
- Response: convolution with the Crystal Ball function

$$\mathrm{CB}(m|\Delta m, \sigma^2, \alpha, \gamma) = \begin{cases} Ce^{-\frac{(m-\Delta m)^2}{2\sigma^2}}, & \frac{m-\Delta m}{\sigma} > -\alpha, \\ C\left(\frac{\gamma}{\alpha}\right)^{\gamma} e^{-\frac{\alpha^2}{2}} \left(\frac{\gamma}{\alpha} - \alpha - \frac{m-\Delta m}{\sigma}\right)^{-\gamma}, & \frac{m-\Delta m}{\sigma} \leq -\alpha \end{cases}$$

- CB parameters estimated with maximum likelihood using 30 % of the data assuming that the true intensity is the non-relativistic Breit–Wigner with PDG values for the Z mass and width
  - Only the remaining 70 % used for unfolding

$Z \to e^+e^-$: Event Display



$Z \to e^+e^-$: Empirical Bayes Unfolding

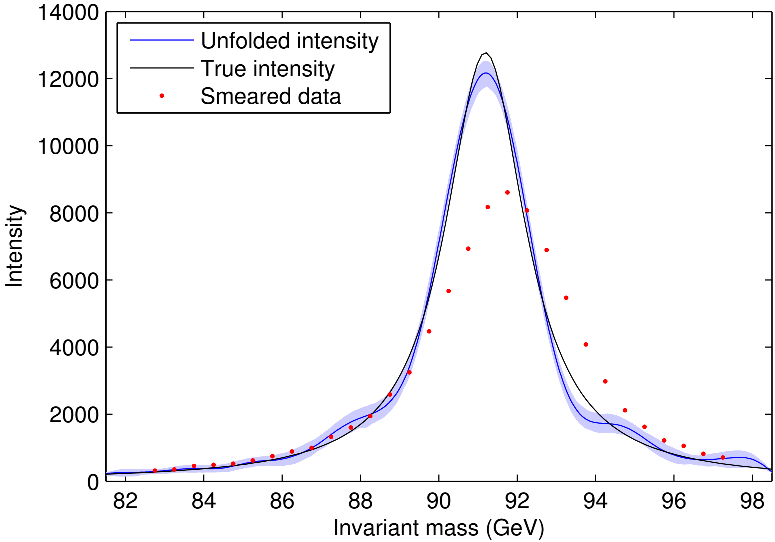


Figure : Empirical Bayes unfolding with bias correction and 95 % pointwise basic intervals

Conclusions

- We have introduced an empirical Bayes unfolding framework which enables a principled choice of the regularization parameter and frequentist uncertainty quantification
- Our studies are motivated by a real-world data analysis problem at CERN
  - We work in direct collaboration with CERN physicists to improve the unfolding techniques used in LHC data analysis
- Our method provides reasonable estimates in very challenging unfolding scenarios
- Uncertainty quantification in unfolding is hampered by the presence of an unavoidable bias from the regularization
  - But basic bootstrap resampling still provides an encouraging first approximation
- Further details in:

Kuusela, M. and Panaretos, V. M. (2014). Empirical Bayes unfolding of elementary particle spectra at the Large Hadron Collider. arXiv:1401.8274 [stat.AP].

References

Chatrchyan, S. et al. (CMS Collaboration, 2013). Energy calibration and resolution of the CMS electromagnetic calorimeter in *pp* collisions at $\sqrt{s} = 7$ TeV. *Journal of Instrumentation*, 8(09):P09009.

Geman, S. and McClure, D. E. (1985). Bayesian image analysis: an application to single photon emission tomography. In *Proceedings of the American Statistical Association, Statistical Computing Section*, pages 12–18.

Geman, S. and McClure, D. E. (1987). Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, LII(4):5–21.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Saquib, S. S., Bouman, C. A., and Sauer, K. (1998). ML parameter estimation for Markov random fields with applications to Bayesian tomography. *IEEE Transactions on Image Processing*, 7(7):1029–1044.

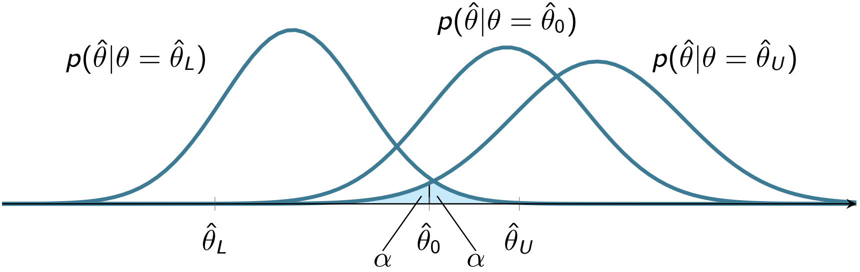
Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:3–36.

Backup

Intuition on the Basic Bootstrap Intervals

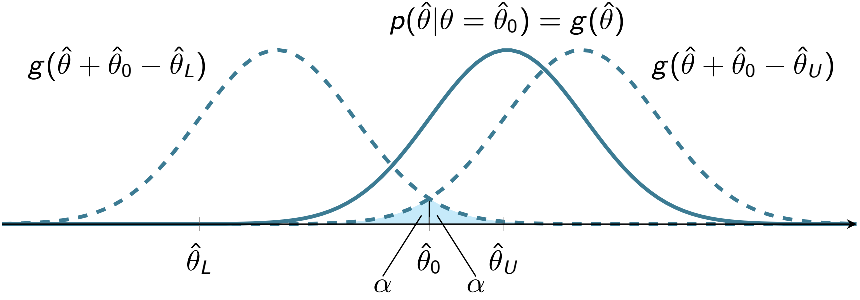
Inversion of a CDF pivot ("Neyman construction"):

$$[\hat{\theta}_L, \hat{\theta}_U] \quad \text{s.t.} \quad \int_{-\infty}^{\hat{\theta}_0} p(\hat{\theta}|\theta = \hat{\theta}_U)\,\mathrm{d}\hat{\theta} = \alpha, \quad \int_{-\infty}^{\hat{\theta}_0} p(\hat{\theta}|\theta = \hat{\theta}_L)\,\mathrm{d}\hat{\theta} = 1 - \alpha$$



Intuition on the Basic Bootstrap Intervals

Basic bootstrap interval: $[\hat{\theta}_L, \hat{\theta}_U] = [2\hat{\theta}_0 - \hat{\theta}^*_{1-\alpha}, 2\hat{\theta}_0 - \hat{\theta}^*_\alpha]$



Demonstration: Setup

λ_{tot}	1 000	20 000
MCEM iterations	30	20
$\delta^{(0)}$	$1 \cdot 10^{-5}$	
MCMC sample size during EM	1 000	500
MCMC sample size for $\hat{\beta}$	1 000	
R	200	
Running time for $\hat{f}$	9 min	3 min
Running time with bootstrap	9 h 56 min	3 h 36 min

$Z \to e^+e^-$: Setup

- We unfold the $n = 30$ bins on the interval $F = [82.5 \text{ GeV}, 97.5 \text{ GeV}]$ and use $p = 38$ B-spline basis functions to reconstruct the true intensity on the interval $E = [81.5 \text{ GeV}, 98.5 \text{ GeV}]$
  - Here $p > n$ facilitates the mixing of the MCMC sampler and $E \supsetneq F$ accounts for boundary effects
- Other parameters:

MCEM iterations	20
$\delta^{(0)}$	$1 \cdot 10^{-6}$
MCMC sample size during EM	500
MCMC sample size for $\hat{\boldsymbol{\beta}}$	5 000
R	200
Running time for $\hat{f}$	5 min
Running time with bootstrap	6 h 13 min

Aristotelian Boundary Conditions (1)

- The prior $p(\boldsymbol{\beta}|\delta) \propto \exp\left(-\delta\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Omega}\boldsymbol{\beta}\right)$ with $\Omega_{i,j} = \int_E B_i''(s)B_j''(s)\,\mathrm{d}s$ is potentially improper since $\boldsymbol{\Omega}$ has rank $p-2$
  - If the prior is improper, then the marginal $p(\mathbf{y}|\delta)$ is also improper and it makes no sense to use empirical Bayes for estimating δ
- The problem can be solved by imposing the so called *Aristotelian boundary conditions*
- That is, we condition on the unknown boundary values of f (or equivalently on β_1 and β_p) and place additional hyperpriors on these values:

$$p(\boldsymbol{\beta}|\delta) = p(\beta_2,\ldots,\beta_{p-1}|\beta_1,\beta_p,\delta)p(\beta_1|\delta)p(\beta_p|\delta), \quad \boldsymbol{\beta} \in \mathbb{R}_+^p,$$

with

$$p(\beta_2,\ldots,\beta_{p-1}|\beta_1,\beta_p,\delta) \propto \exp(-\delta\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Omega}\boldsymbol{\beta}),$$
$$p(\beta_1|\delta) \propto \exp\left(-\delta\gamma_{\mathrm{L}}\beta_1^2\right),$$
$$p(\beta_p|\delta) \propto \exp\left(-\delta\gamma_{\mathrm{R}}\beta_p^2\right),$$

where $\gamma_{\mathrm{L}}, \gamma_{\mathrm{R}} > 0$ are fixed constants

Aristotelian Boundary Conditions (2)

- As a result $p(\boldsymbol{\beta}|\delta) \propto \exp\left(-\delta\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Omega}_{\mathrm{A}}\boldsymbol{\beta}\right)$ where the elements of $\boldsymbol{\Omega}_{\mathrm{A}}$ are given by

$$\Omega_{\mathrm{A},i,j} = \begin{cases} \Omega_{i,j} + \gamma_{\mathrm{L}}, & \text{if } i = j = 1, \\ \Omega_{i,j} + \gamma_{\mathrm{R}}, & \text{if } i = j = p, \\ \Omega_{i,j}, & \text{otherwise} \end{cases}$$

- The augmented matrix $\boldsymbol{\Omega}_{\mathrm{A}}$ is positive definite and hence the modified prior is a proper pdf
- The Aristotelian prior has the added benefit that by controlling γ_{L} and γ_{R} we are able to control the variance of $\hat{f}$ near the boundaries
- In our numerical experiments we had:
  - Gaussian mixture model data: $\gamma_{\mathrm{L}} = \gamma_{\mathrm{R}} = 5$
  - $Z \to e^+e^-$ data: $\gamma_{\mathrm{L}} = \gamma_{\mathrm{R}} = 70$

Demonstration: No regularization

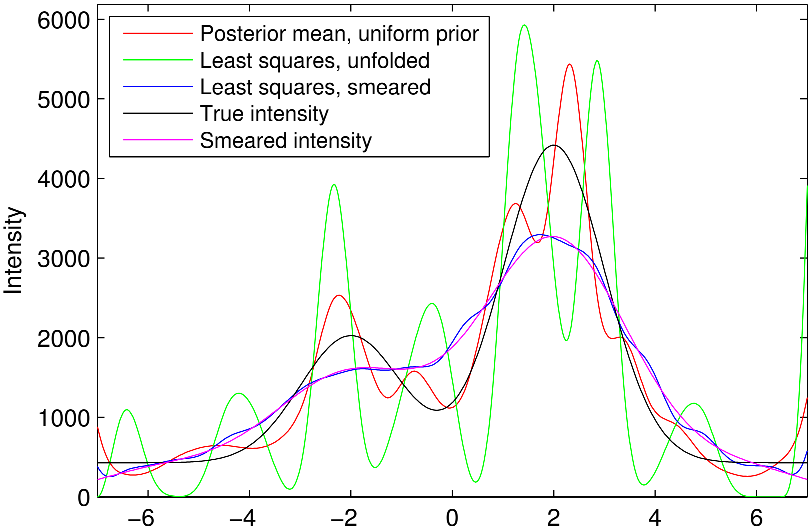


Figure : Unfolding of the Gaussian mixture model data ($\lambda_{\mathrm{tot}} = 20\,000$) *without regularization*.

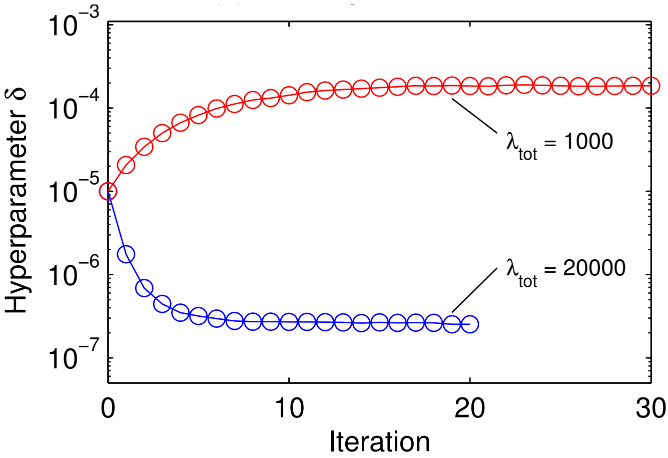


Figure : Convergence of the MCEM algorithm for estimating the hyperparameter δ with the Gaussian mixture model data

MCMC Diagnostics

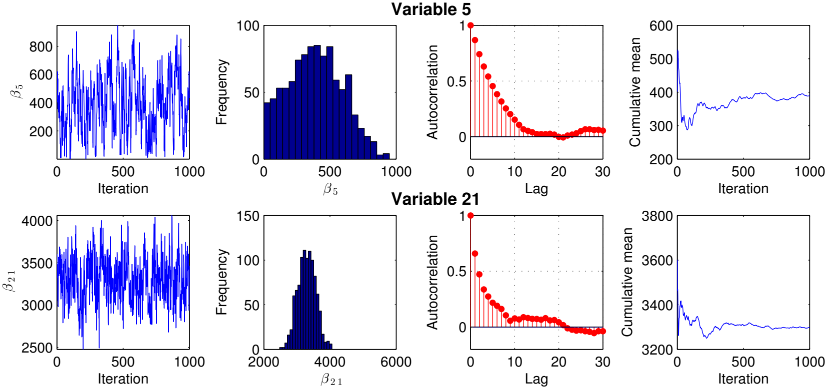


Figure : Convergence and mixing diagnostics for the single-component Metropolis–Hastings sampler for variables β_5 and β_{21} with the Gaussian mixture model data with $\lambda_{\text{tot}} = 20\,000$: from left to right, the trace plots, histograms, estimated autocorrelation functions and cumulative means of the samples.

Convergence of Empirical Bayes Unfolding

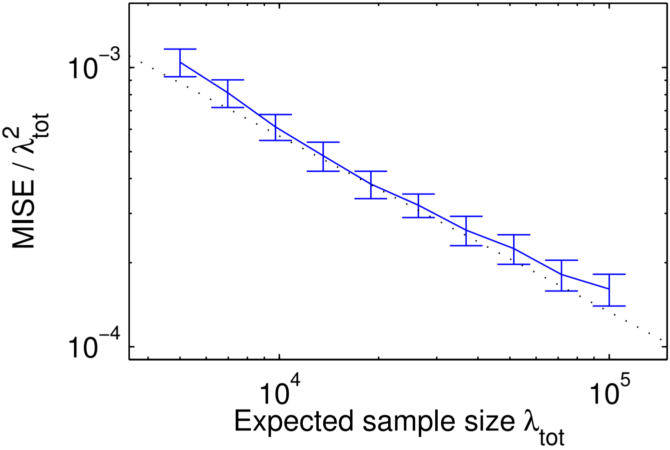


Figure : Convergence of the mean integrated squared error (MISE) with the Gaussian mixture model data as the expected sample size λ_{tot} grows. The error bars indicate approximate 95 % confidence intervals.

Monte Carlo EM Algorithm for Finding the MMLE

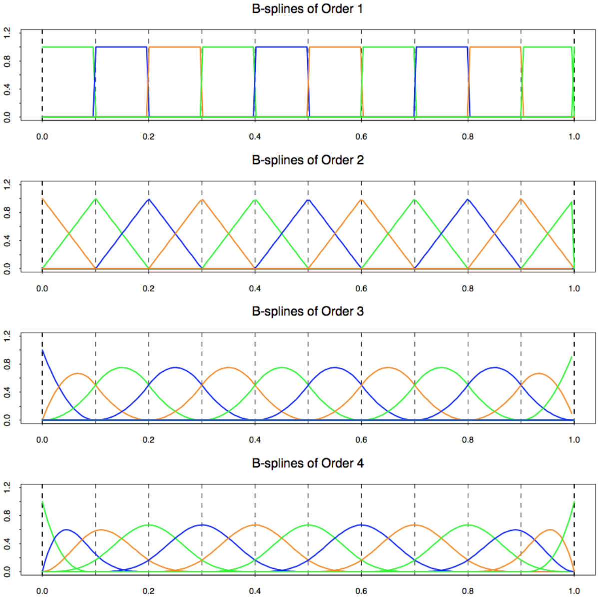
The Monte Carlo EM algorithm (Geman and McClure, 1985, 1987; Saquib et al., 1998) for finding the marginal maximum likelihood estimate $\hat{\delta}$:

1. Pick some initial guess $\delta^{(0)} > 0$ and set $t = 0$
2. E-step:
    1. Sample $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots, \boldsymbol{\beta}^{(S)}$ from the posterior $p(\boldsymbol{\beta}|\mathbf{y}, \delta^{(t)})$
    2. Compute:

$$Q(\delta; \delta^{(t)}) = \frac{1}{S} \sum_{s=1}^{S} \log p(\boldsymbol{\beta}^{(s)}|\delta)$$

3. M-step: Set $\delta^{(t+1)} = \arg\max_{\delta > 0} Q(\delta; \delta^{(t)})$
4. Set $t \leftarrow t + 1$
5. If some stopping rule is satisfied, set $\hat{\delta} = \delta^{(t)}$ and terminate the iteration, else go to step 2

B-Spline Basis Functions



Details of the MCMC Implementation (1)

- We use the single-component Metropolis–Hastings sampler of Saquib et al. (1998)
- The kth full conditional satisfies

$$\log p(\beta_k|\boldsymbol{\beta}_{-k}, \mathbf{y}, \delta) = \sum_{i=1}^{n} y_i \log\left(\sum_{j=1}^{p} K_{i,j}\beta_j\right) - \sum_{i=1}^{n}\sum_{j=1}^{p} K_{i,j}\beta_j - \delta \sum_{i=1}^{p}\sum_{j=1}^{p} \Omega_{i,j}\beta_i\beta_j + \text{const} := f(\beta_k, \boldsymbol{\beta}_{-k})$$

- Taking a 2nd order Taylor expansion of the log-term around the current position β_k of the Markov chain, we find

$$f(\beta_k^*, \boldsymbol{\beta}_{-k}) \approx d_{1,k}(\beta_k^* - \beta_k) + \frac{d_{2,k}}{2}(\beta_k^* - \beta_k)^2 - \delta\left(\Omega_{k,k}(\beta_k^*)^2 + 2\sum_{i \neq k} \Omega_{i,k}\beta_i\beta_k^*\right) + \text{const} := g(\beta_k^*, \boldsymbol{\beta}),$$

where

$$d_{1,k} = -\sum_{i=1}^{n} K_{i,k}\left(1 - \frac{y_i}{\mu_i}\right), \quad d_{2,k} = -\sum_{i=1}^{n} y_i \left(\frac{K_{i,k}}{\mu_i}\right)^2$$

with $\boldsymbol{\mu} = \mathbf{K}\boldsymbol{\beta}$

Details of the MCMC Implementation (2)

- As a function of β_k^*, the approximate full conditional

$$g(\beta_k^*, \boldsymbol{\beta}) = d_{1,k}(\beta_k^* - \beta_k) + \frac{d_{2,k}}{2}(\beta_k^* - \beta_k)^2 - \delta\left(\Omega_{k,k}(\beta_k^*)^2 + 2\sum_{i\neq k}\Omega_{i,k}\beta_i\beta_k^*\right) + \mathrm{const}$$

  is a Gaussian with mean

$$m_k = \frac{d_{1,k} - d_{2,k}\beta_k - 2\delta\sum_{i\neq k}\Omega_{i,k}\beta_i}{2\delta\Omega_{k,k} - d_{2,k}}$$

  and variance

$$\sigma_k^2 = \frac{1}{2\delta\Omega_{k,k} - d_{2,k}}$$

Details of the MCMC Implementation (3)

- If $m_k \geq 0$, the proposal β_k^* is sampled from $\mathcal{N}(m_k, \sigma_k^2)$ truncated to $[0, \infty)$
- If $m_k < 0$, the proposal β_k^* is sampled from $\mathrm{Exp}(\lambda)$ with

$$\frac{\partial}{\partial \beta_k^*} \log p(\beta_k^* | \boldsymbol{\beta}) \bigg|_{\beta_k^*=0} = \frac{\partial}{\partial \beta_k^*} g(\beta_k^*, \boldsymbol{\beta}) \bigg|_{\beta_k^*=0}$$

  giving $\lambda = -d_{1,k} + d_{2,k}\beta_k + 2\delta \sum_{i \neq k} \Omega_{i,k}\beta_i$
- Denote: $p(\beta_k^* | \boldsymbol{\beta}) := q(\beta_k^*, \beta_k, \boldsymbol{\beta}_{-k})$, $p(\boldsymbol{\beta} | \mathbf{y}, \delta) := h(\beta_k, \boldsymbol{\beta}_{-k})$
- The acceptance probability for the kth component of the single-component Metropolis–Hastings algorithm is given by

$$a(\beta_k^*, \boldsymbol{\beta}) = \min \left\{ 1, \frac{h(\beta_k^*, \boldsymbol{\beta}_{-k}) q(\beta_k, \beta_k^*, \boldsymbol{\beta}_{-k})}{h(\beta_k, \boldsymbol{\beta}_{-k}) q(\beta_k^*, \beta_k, \boldsymbol{\beta}_{-k})} \right\}$$

The Expectation-Maximization Algorithm

- The *EM algorithm* is an iterative method for finding the maximum of the likelihood $L(\boldsymbol{\theta}; \mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})$
- Applies in cases where the data $\mathbf{y}$ can be seen as an incomplete version of some complete data $\mathbf{x}$ (that is, $\mathbf{y} = g(\mathbf{x})$) with complete-data likelihood $L(\boldsymbol{\theta}; \mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta})$
- The EM iteration:
  1. Pick some initial guess $\boldsymbol{\theta}^{(0)}$ and set $t = 0$
  2. E-step: Compute $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = \mathrm{E}(\log p(\mathbf{x}|\boldsymbol{\theta})|\mathbf{y}, \boldsymbol{\theta}^{(t)})$
  3. M-step: Set $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$
  4. Set $t \leftarrow t + 1$
  5. If some stopping rule is satisfied, set $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \boldsymbol{\theta}^{(t)}$ and terminate the iteration, else go to step 2
- The EM iteration never decreases the incomplete-data likelihood
  - That is, $L(\boldsymbol{\theta}^{(t+1)}; \mathbf{y}) \geq L(\boldsymbol{\theta}^{(t)}; \mathbf{y})$ for all $t = 0, 1, 2, \ldots$