Structured Convex Optimization over Probability Measures

Yeongcheong Baek, Chris Jordan-Squire
Jim Burke
Laboratory for Applied Pharmacokinetics and Bioinformatics
Childrens Hospital of Los Angeles

Workshop on Optimization and Learning: theory and applications
Centre de Recherches Matématiques, Montréal, May 30, 2025

Outline

I. Motivating example: Pharmacokenetic clearance

II. The general problem $\min_{\mu \in \mathcal{P}(\Omega)} \psi(S\mu)$

III. Duality Theory

IV. Reduction to finite dimensional convex-composite optimization

V. Algorithms

Pharmacokenetic clearance

A simple one compartment model of drug clearance from plasma.

Pharmacokenetic clearance

A simple one compartment model of drug clearance from plasma. After an injection, the concentration of a drug remaining in the body at time t is modeled by exponential decay:

$$g(t, \beta) := \frac{\exp(-K\,t)}{V}.$$

The unknown parameters to be estimated are

$$\beta := \begin{bmatrix} K \\ V \end{bmatrix} = \begin{bmatrix} \text{clearance rate} \\ \text{blood volume} \end{bmatrix}.$$

Pharmacokenetic clearance

A simple one compartment model of drug clearance from plasma. After an injection, the concentration of a drug remaining in the body at time t is modeled by exponential decay:

$$g(t, \beta) := \frac{\exp(-K\,t)}{V}.$$

The unknown parameters to be estimated are

$$\beta := \begin{bmatrix} K \\ V \end{bmatrix} = \begin{bmatrix} \text{clearance rate} \\ \text{blood volume} \end{bmatrix}.$$

The observations are blood draws at times $t_1, \ldots, t_N$,

$$y = G(\beta) + \epsilon = \begin{pmatrix} g(t_1, \beta) + \epsilon_1 \\ \vdots \\ g(t_N, \beta) + \epsilon_N \end{pmatrix}.$$

Pharmacokenetic clearance

A simple one compartment model of drug clearance from plasma. After an injection, the concentration of a drug remaining in the body at time t is modeled by exponential decay:

$$g(t, \beta) := \frac{\exp(-K\,t)}{V}.$$

The unknown parameters to be estimated are

$$\beta := \begin{bmatrix} K \\ V \end{bmatrix} = \begin{bmatrix} \text{clearance rate} \\ \text{blood volume} \end{bmatrix}.$$

The observations are blood draws at times $t_1, \ldots, t_N$,

$$y = G(\beta) + \epsilon = \begin{pmatrix} g(t_1, \beta) + \epsilon_1 \\ \vdots \\ g(t_N, \beta) + \epsilon_N \end{pmatrix}.$$

The observation errors, ϵ_j, are depend on the individual from which the sample is taken.

Model error for the individual

We assume that the model error for the individual comes from a known parametric family of densities $P(\epsilon \mid \beta)$. For example, a normal family $\mathcal{N}(0, R(\beta))$ so that

$$P(y \mid \beta) = \left(\frac{1}{|2\pi R(\beta)|}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2}(y - G(\beta))^T R(\beta)^{-1}(y - G(\beta))\right].$$

Model error for the individual

We assume that the model error for the individual comes from a known parametric family of densities $P(\epsilon \mid \beta)$. For example, a normal family $\mathcal{N}(0, R(\beta))$ so that

$$P(y \mid \beta) = \left(\frac{1}{|2\pi R(\beta)|}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2}(y - G(\beta))^T R(\beta)^{-1}(y - G(\beta))\right].$$

For example in a clearance model, one often takes something similar to

$$y_j \sim \mathcal{N}(g(t_j, \beta), g(t_j, \beta)^2),\ j = 1, \ldots, N.$$

Model error for the individual

We assume that the model error for the individual comes from a known parametric family of densities $P(\epsilon \mid \beta)$. For example, a normal family $\mathcal{N}(0, R(\beta))$ so that

$$P(y \mid \beta) = \left(\frac{1}{|2\pi R(\beta)|}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2}(y - G(\beta))^T R(\beta)^{-1}(y - G(\beta))\right].$$

For example in a clearance model, one often takes something similar to

$$y_j \sim \mathcal{N}(g(t_j, \beta), g(t_j, \beta)^2),\ j = 1, \dots, N.$$

In a population study, one studies a population of individuals (think phase 3 clinical trials). That is, we have many observations from distinct individuals $y^i \in \mathbb{R}^N,\ i = 1, \dots, m$, however, N, the number of samples per individual is small.

Nonparametric population density estimation

- Goal: Estimate $P(y \mid \mu)$ for $\mu \in \mathcal{P}(\Omega)$ over the population.

Nonparametric population density estimation

- Goal: Estimate $P(y \mid \mu)$ for $\mu \in \mathcal{P}(\Omega)$ over the population.

- Assumption: We can estimate the measure μ as a mixture of the individual error models:

$$P(y \mid \mu) = \int_{\Omega} P(y \mid \beta) d\mu(\beta).$$

Nonparametric population density estimation

- Goal: Estimate $P(y \mid \mu)$ for $\mu \in \mathcal{P}(\Omega)$ over the population.

- Assumption: We can estimate the measure μ as a mixture of the individual error models:

$$P(y \mid \mu) = \int_{\Omega} P(y \mid \beta) d\mu(\beta).$$

- Individualized medicine:

The *nonparametric* setting is chosen since we wish to discover subpopulations (or modes) within the population having distinctly different clearance profiles.

This population estimate can be used for covariate discovery to explain the variability. It can also be used as a prior to quickly identify the therapeutic treatment range for a new patients.

Sample Results: Single plasma bolus

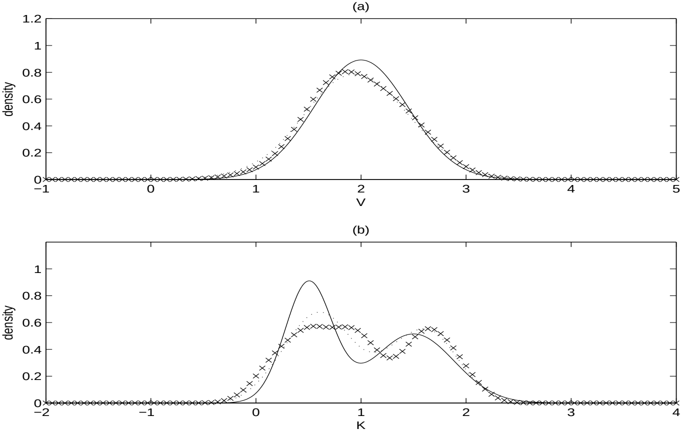

Figure: Marginal probability density functions of (a) V and (b) K ; solid line, true density; dotted line, smoothed sample distribution; x–line, smoothed optimal solution

Nonparametric Maximum Likelihood (NPML)

Given observation $y^1, \ldots, y^m \in \mathbb{R}^N$ solve

$$\min_{\mu \in \mathcal{P}(\Omega)} L(\mu) := \varphi\left(\int_\Omega F(\beta)\mu(d\beta)\right) + \delta_K\left(\int_\Omega H(\beta)\mu(d\beta)\right),$$

where $\Omega \subset \mathbb{R}^n$ compact, $K \subset \mathbb{R}^s$ is closed convex,

$$F(\beta) = \begin{pmatrix} P(y^1|\beta) \\ \vdots \\ P(y^m|\beta) \end{pmatrix} \text{ and } \varphi(z) = \begin{cases} -\sum_{i=1}^m \log(z_i) & , \ z \in \mathbb{R}^m_{++}, \\ +\infty & , \ \text{else.} \end{cases}$$

Nonparametric Maximum Likelihood (NPML)

Given observation $y^1, \ldots, y^m \in \mathbb{R}^N$ solve

$$\min_{\mu \in \mathcal{P}(\Omega)} L(\mu) := \varphi\left(\int_\Omega F(\beta)\mu(d\beta)\right) + \delta_K\left(\int_\Omega H(\beta)\mu(d\beta)\right),$$

where $\Omega \subset \mathbb{R}^n$ compact, $K \subset \mathbb{R}^s$ is closed convex,

$$F(\beta) = \begin{pmatrix} P(y^1|\beta) \\ \vdots \\ P(y^m|\beta) \end{pmatrix} \text{ and } \varphi(z) = \begin{cases} -\sum_{i=1}^m \log(z_i) & , \ z \in \mathbb{R}^m_{++}, \\ +\infty & , \text{ else.} \end{cases}$$

Examples of component functions for $H : \mathbb{R}^n \to \mathbb{R}^s$:

(i) Moment constraints, e.g. the mean $\int_\Omega \beta\mu(d\beta) = \theta$.

(ii) Mean-Variance constraints:

$$\int_\Omega \beta\mu(d\beta) = \theta, \qquad \Sigma_l \preceq \int_\Omega (\beta-\theta)(\beta-\theta)^T \mu(d\beta) \preceq \Sigma_u \ .$$

NPML, the unconstrained case

$$(\mathtt{P})_{\mathsf{NPML}} \qquad \min_{\mu \in \mathcal{P}(\Omega)} \varphi\left(\int_\Omega F(\beta)\mu(d\beta)\right)$$

NPML, the unconstrained case

$$(\mathtt{P})_{\mathsf{NPML}} \qquad \min_{\mu \in \mathcal{P}(\Omega)} \varphi\left(\int_{\Omega} F(\beta)\mu(d\beta)\right)$$

$\mu \mapsto \int_{\Omega} F(\beta)\mu(d\beta)$ is a continuous linear transformation $\mathcal{B}(\Omega) \mapsto \mathbb{R}^m$.
Denote this linear mapping by $S \in \mathcal{L}[\mathcal{B}(\Omega), \mathbb{R}^m]$, that is

$$\varphi(S\mu) = \varphi\left(\int_{\Omega} F(\beta)\mu(d\beta)\right) \text{ with } F(\beta) = \begin{bmatrix} P(y^1|\beta) \\ \vdots \\ P(y^m|\beta) \end{bmatrix}.$$

NPML, the unconstrained case

$$(\mathtt{P})_{\mathsf{NPML}} \qquad \min_{\mu\in\mathcal{P}(\Omega)} \varphi\left(\int_\Omega F(\beta)\mu(d\beta)\right)$$

$\mu \mapsto \int_\Omega F(\beta)\mu(d\beta)$ is a continuous linear transformation $\mathcal{B}(\Omega) \mapsto \mathbb{R}^m$. Denote this linear mapping by $S \in \mathcal{L}[\mathcal{B}(\Omega), \mathbb{R}^m]$, that is

$$\varphi(S\mu) = \varphi\left(\int_\Omega F(\beta)\mu(d\beta)\right) \text{ with } F(\beta) = \begin{bmatrix} P(y^1|\beta) \\ \vdots \\ P(y^m|\beta) \end{bmatrix}.$$

Therefore, $(\mathtt{P})$ can be written as

$$\min_{\mu\in\mathcal{P}(\Omega)} \varphi(S\mu) = \min_{w\in C} \varphi(w),$$

where $C := S[\mathcal{P}(\Omega)]$ is the linear image of the w*-compact convex set $\mathcal{P}(\Omega)$ and so C is compact convex.

NPML, the unconstrained case

$$(\texttt{P})_{\textsf{NPML}} \qquad \min_{\mu\in\mathcal{P}(\Omega)} \varphi\left(\int_\Omega F(\beta)\mu(d\beta)\right)$$

$\mu \mapsto \int_\Omega F(\beta)\mu(d\beta)$ is a continuous linear transformation $\mathcal{B}(\Omega) \mapsto \mathbb{R}^m$. Denote this linear mapping by $S \in \mathcal{L}[\mathcal{B}(\Omega), \mathbb{R}^m]$, that is

$$\varphi(S\mu) = \varphi\left(\int_\Omega F(\beta)\mu(d\beta)\right) \text{ with } F(\beta) = \begin{bmatrix} P(y^1|\beta) \\ \vdots \\ P(y^m|\beta) \end{bmatrix}.$$

Therefore, $(\texttt{P})$ can be written as

$$\min_{\mu\in\mathcal{P}(\Omega)} \varphi(S\mu) = \min_{w\in C} \varphi(w),$$

where $C := S[\mathcal{P}(\Omega)]$ is the linear image of the w*-compact convex set $\mathcal{P}(\Omega)$ and so C is compact convex.

$$\boxed{(\mathrm{CQ})_{\textsf{NPML}} \qquad \exists\ \beta \in \Omega \text{ s.t. } P(y^i|\beta) > 0,\ i = 1, \ldots, m.}$$

NPML, the unconstrained case

$$(\mathtt{P})_{\mathsf{NPML}} \qquad \min_{\mu \in \mathcal{P}(\Omega)} \varphi\left(\int_\Omega F(\beta)\mu(d\beta)\right)$$

$\mu \mapsto \int_\Omega F(\beta)\mu(d\beta)$ is a continuous linear transformation $\mathcal{B}(\Omega) \mapsto \mathbb{R}^m$. Denote this linear mapping by $S \in \mathcal{L}[\mathcal{B}(\Omega), \mathbb{R}^m]$, that is

$$\varphi(S\mu) = \varphi\left(\int_\Omega F(\beta)\mu(d\beta)\right) \text{ with } F(\beta) = \begin{bmatrix} P(y^1|\beta) \\ \vdots \\ P(y^m|\beta) \end{bmatrix}.$$

Therefore, $(\mathtt{P})$ can be written as

$$\min_{\mu \in \mathcal{P}(\Omega)} \varphi(S\mu) = \min_{w \in C} \varphi(w),$$

where $C := S[\mathcal{P}(\Omega)]$ is the linear image of the w*-compact convex set $\mathcal{P}(\Omega)$ and so C is compact convex.

$$\boxed{(\mathrm{CQ})_{\mathsf{NPML}} \qquad \exists\ \beta \in \Omega \text{ s.t. } P(y^i|\beta) > 0,\ i = 1, \ldots, m.}$$

$(\mathrm{CQ})_{\mathsf{NPML}}$ implies strong duality.

The Full Problem Class

$$(\texttt{P}) \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}(\Omega)}(\mu).$$

The Full Problem Class

$$(\mathtt{P}) \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}(\Omega)}(\mu).$$

$\Omega \subset \mathbb{R}^n$ compact (e.g. a big box)

The Full Problem Class

$$(\mathtt{P}) \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}(\Omega)}(\mu).$$

$\Omega \subset \mathbb{R}^n$ compact (e.g. a big box)

$C(\Omega)$ continuous functions on Ω (sup-norm)

The Full Problem Class

$$\text{(P)} \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}(\Omega)}(\mu).$$

$\Omega \subset \mathbb{R}^n$ compact (e.g. a big box)

$C(\Omega)$ continuous functions on Ω (sup-norm)

$\mathcal{B}(\Omega)$ Borel measures over Ω $\qquad$ ($C(\Omega)^* = \mathcal{B}(\Omega)$)

The Full Problem Class

$$(\mathtt{P}) \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}(\Omega)}(\mu).$$

$\Omega \subset \mathbb{R}^n$ compact (e.g. a big box)

$C(\Omega)$ continuous functions on Ω (sup-norm)

$\mathcal{B}(\Omega)$ Borel measures over Ω $\qquad (C(\Omega)^* = \mathcal{B}(\Omega))$

$\psi : \mathbb{R}^m \to \mathbb{R}_{\mathsf{e}} := \mathbb{R} \cup \{+\infty\}$ lsc, proper, convex

The Full Problem Class

$$\text{(P)} \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}(\Omega)}(\mu).$$

$\Omega \subset \mathbb{R}^n$ compact (e.g. a big box)

$C(\Omega)$ continuous functions on Ω (sup-norm)

$\mathcal{B}(\Omega)$ Borel measures over Ω $\qquad (C(\Omega)^* = \mathcal{B}(\Omega))$

$\psi : \mathbb{R}^m \to \mathbb{R}_{\mathsf{e}} := \mathbb{R} \cup \{+\infty\}$ lsc, proper, convex

$S \in \mathcal{L}(\mathcal{B}(\Omega), \mathbb{R}^m)$ continuous linear transformation

The Full Problem Class

$$\text{(P)} \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}(\Omega)}(\mu).$$

$\Omega \subset \mathbb{R}^n$ compact (e.g. a big box)

$C(\Omega)$ continuous functions on Ω (sup-norm)

$\mathcal{B}(\Omega)$ Borel measures over Ω $\qquad (C(\Omega)^* = \mathcal{B}(\Omega))$

$\psi : \mathbb{R}^m \to \mathbb{R}_\mathsf{e} := \mathbb{R} \cup \{+\infty\}$ lsc, proper, convex

$S \in \mathcal{L}(\mathcal{B}(\Omega), \mathbb{R}^m)$ continuous linear transformation

$\mathcal{P}(\Omega) \subset \mathcal{B}(\Omega)$ probability measures (convex w*-compact set)

The Full Problem Class

$$(\mathtt{P}) \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}(\Omega)}(\mu).$$

$\Omega \subset \mathbb{R}^n$ compact (e.g. a big box)

$C(\Omega)$ continuous functions on Ω (sup-norm)

$\mathcal{B}(\Omega)$ Borel measures over Ω $\qquad (C(\Omega)^* = \mathcal{B}(\Omega))$

$\psi : \mathbb{R}^m \to \mathbb{R}_{\mathsf{e}} := \mathbb{R} \cup \{+\infty\}$ lsc, proper, convex

$S \in \mathcal{L}(\mathcal{B}(\Omega), \mathbb{R}^m)$ continuous linear transformation

$\mathcal{P}(\Omega) \subset \mathcal{B}(\Omega)$ probability measures (convex w*-compact set)

$$\delta_{\mathcal{P}(\Omega)}(\mu) := \begin{cases} 0 & , \ \mu \in \mathcal{P}(\Omega), \\ +\infty & , \ \mu \notin \mathcal{P}(\Omega). \end{cases}$$

Applications

1. nonparametric mixture models (Lindsay 83' and 95' book)
    1. mixed effects models (98'-, USC Pharmacokinetics Lab)
    2. repeated measure models
    3. latent class models
    4. missing data models
    5. nuisance parameter models
    6. deconvolution models
    7. clustering
2. optimal experimental design
   (Fedorov 72' book)
3. maximum entropy problems
   (Berger-Pietra-Pietra 96' for Nat. Language Processing)
4. distributionally robust stochastic programming
   (Shapiro-Kleywergt 2002)

Duality for (P) $\min_{\mu\in\mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}(\Omega)}(\mu)$.

• $\mathcal{B}(\Omega)$ (w*-topology) and $C(\Omega)$ (sup-norm topology) are paired in duality via the pairing

$$\langle \mu,\, f\rangle := \int_\Omega f(\omega)d\mu(\omega) \quad \forall\,(\mu, f) \in \mathcal{B}(\Omega)\times C(\Omega).$$

Recall, $C(\Omega)^* = \mathcal{B}(\Omega)$ but $\mathcal{B}(\Omega)^* \neq C(\Omega)$.

Duality for (P) $\min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}(\Omega)}(\mu)$.

• $\mathcal{B}(\Omega)$ (w*-topology) and $C(\Omega)$ (sup-norm topology) are paired in duality via the pairing

$$\langle \mu,\, f \rangle := \int_\Omega f(\omega) d\mu(\omega) \quad \forall\, (\mu, f) \in \mathcal{B}(\Omega) \times C(\Omega).$$

Recall, $C(\Omega)^* = \mathcal{B}(\Omega)$ but $\mathcal{B}(\Omega)^* \neq C(\Omega)$.

• For $g : \mathcal{B}(\Omega) \to \mathbb{R}_\mathsf{e}$, the convex conjugate of g is given by

$$g^*(\phi) := \sup_{\mu \in \mathcal{B}(\Omega)} [\langle \mu,\, \phi \rangle - g(\mu)] \qquad \forall\, \phi \in C(\Omega),$$

where $g^* : C(\Omega) \to \mathbb{R}_\mathsf{e}$.

Duality for (P) $\min_{\mu\in\mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}(\Omega)}(\mu)$.

• $\mathcal{B}(\Omega)$ (w*-topology) and $C(\Omega)$ (sup-norm topology) are paired in duality via the pairing

$$\langle \mu,\, f\rangle := \int_\Omega f(\omega)d\mu(\omega) \quad \forall\,(\mu, f) \in \mathcal{B}(\Omega) \times C(\Omega).$$

Recall, $C(\Omega)^* = \mathcal{B}(\Omega)$ but $\mathcal{B}(\Omega)^* \neq C(\Omega)$.

• For $g\,:\,\mathcal{B}(\Omega) \to \mathbb{R}_e$, the convex conjugate of g is given by

$$g^*(\phi) := \sup_{\mu\in\mathcal{B}(\Omega)} [\langle \mu,\, \phi\rangle - g(\mu)] \qquad \forall\,\phi \in C(\Omega),$$

where $g^*\,:\,C(\Omega) \to \mathbb{R}_e$.

Example: The support function for $\mathcal{P}(\Omega)$ is the conjugate of $\delta_{\mathcal{P}(\Omega)}$: for all $f \in C(\Omega)$,

$$\delta^*_{\mathcal{P}(\Omega)}(f) = \sup_{\mu\in\mathcal{P}(\Omega)} \langle \mu,\, f\rangle = \sup_{\mu\in\mathcal{P}(\Omega)} \int_\Omega f(\beta)d\mu(\beta) = \max_{\beta\in\Omega} f(\beta).$$

Duality for (P) $\min_{\mu\in\mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}(\Omega)}(\mu)$.

• $\mathcal{B}(\Omega)$ (w*-topology) and $C(\Omega)$ (sup-norm topology) are paired in duality via the pairing

$$\langle \mu,\, f\rangle := \int_\Omega f(\omega)d\mu(\omega) \quad \forall\,(\mu, f) \in \mathcal{B}(\Omega)\times C(\Omega).$$

Recall, $C(\Omega)^* = \mathcal{B}(\Omega)$ but $\mathcal{B}(\Omega)^* \neq C(\Omega)$.

• For $g : \mathcal{B}(\Omega) \to \mathbb{R}_e$, the convex conjugate of g is given by

$$g^*(\phi) := \sup_{\mu\in\mathcal{B}(\Omega)} [\langle\mu,\, \phi\rangle - g(\mu)] \qquad \forall\,\phi \in C(\Omega),$$

where $g^* : C(\Omega) \to \mathbb{R}_e$.

Example: The support function for $\mathcal{P}(\Omega)$ is the conjugate of $\delta_{\mathcal{P}(\Omega)}$: for all $f \in C(\Omega)$,

$$\delta^*_{\mathcal{P}(\Omega)}(f) = \sup_{\mu\in\mathcal{P}(\Omega)} \langle\mu,\, f\rangle = \sup_{\mu\in\mathcal{P}(\Omega)} \int_\Omega f(\beta)d\mu(\beta) = \max_{\beta\in\Omega} f(\beta).$$

Dual Convex Programs

$$\texttt{(P)} \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}}(\mu)$$

$$\texttt{(D)} \qquad \min_{w \in \mathbb{R}^m} \psi^*(-w) + \delta^*_{\mathcal{P}}(S^* w)$$

Dual Convex Programs

$$(\texttt{P}) \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}}(\mu)$$

$$(\texttt{D}) \qquad \min_{w \in \mathbb{R}^m} \psi^*(-w) + \delta_{\mathcal{P}}^*(S^* w)$$

where $S^* \in \mathcal{L}[\mathbb{R}^m, C(\Omega)]$ is the unique solution to

$$\langle w,\, S\mu \rangle_{\mathbb{R}^m} = \langle S^* w,\, \mu \rangle \quad \forall\, (w, \mu) \in \mathbb{R}^m \times \mathcal{B}(\Omega).$$

Dual Convex Programs

$$(\texttt{P}) \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}}(\mu)$$

$$(\texttt{D}) \qquad \min_{w \in \mathbb{R}^m} \psi^*(-w) + \delta_{\mathcal{P}}^*(S^* w)$$

where $S^* \in \mathcal{L}[\mathbb{R}^m, C(\Omega)]$ is the unique solution to

$$\langle w,\, S\mu \rangle_{\mathbb{R}^m} = \langle S^* w,\, \mu \rangle \quad \forall\, (w, \mu) \in \mathbb{R}^m \times \mathcal{B}(\Omega).$$

Riesz representation theorem implies there exists a continuous mapping $F : \Omega \to \mathbb{R}^m$ such that

$$S\mu = \int_\Omega F(\beta) d\mu \quad \text{and} \quad S^* w = \langle w,\, F \rangle_{\mathbb{R}^m} \in C(\Omega),$$

where $\langle w,\, F \rangle_{\mathbb{R}^m} (\beta) := \langle w,\, F(\beta) \rangle$. Hence

$$\delta_{\mathcal{P}}^*(S^* w) = \sup_{\beta \in \Omega} \langle w,\, F(\beta) \rangle\,.$$

Duality: NPML, the unconstrained case

$$(\mathtt{P})_{\mathsf{NPML}} \qquad \min_{\mu \in \mathcal{P}(\Omega)} \phi\left(\int_\Omega F(\beta)\mu(d\beta)\right)$$

Duality: NPML, the unconstrained case

$$(\mathtt{P})_{\mathsf{NPML}} \qquad \min_{\mu \in \mathcal{P}(\Omega)} \phi\left(\int_\Omega F(\beta)\mu(d\beta)\right)$$

$$(\mathtt{D})_{\mathsf{NPML}} \qquad \min_{w \in \mathbb{R}^m} \left[\phi^*(-w) + \sup_{\beta \in \Omega} \langle w,\, F(\beta)\rangle\right]$$

Duality: NPML, the unconstrained case

$$(\mathtt{P})_{\mathsf{NPML}} \qquad \min_{\mu\in\mathcal{P}(\Omega)} \phi\left(\int_\Omega F(\beta)\mu(d\beta)\right)$$

$$(\mathtt{D})_{\mathsf{NPML}} \qquad \min_{w\in\mathbb{R}^m} \left[\phi^*(-w) + \sup_{\beta\in\Omega} \langle w,\, F(\beta)\rangle\right]$$

Here, $\langle w,\, F(\beta)\rangle = \sum_{i=1}^m w_i P(y_i|\beta)$.

Duality: NPML, the unconstrained case

$$(\mathtt{P})_{\mathsf{NPML}} \qquad \min_{\mu \in \mathcal{P}(\Omega)} \phi\left(\int_{\Omega} F(\beta)\mu(d\beta)\right)$$

$$(\mathtt{D})_{\mathsf{NPML}} \qquad \min_{w \in \mathbb{R}^m} \left[\phi^*(-w) + \sup_{\beta \in \Omega} \langle w,\, F(\beta)\rangle\right]$$

Here, $\langle w,\, F(\beta)\rangle = \sum_{i=1}^{m} w_i P(y_i|\beta)$.

$$(\mathrm{CQ}) \qquad \exists\ \beta \in \Omega \text{ s.t. } P(y_i|\beta) > 0,\ i = 1, \ldots, m.$$

Duality: NPML, the unconstrained case

$$(\mathtt{P})_{\mathsf{NPML}} \qquad \min_{\mu \in \mathcal{P}(\Omega)} \phi\left(\int_\Omega F(\beta)\mu(d\beta)\right)$$

$$(\mathtt{D})_{\mathsf{NPML}} \qquad \min_{w \in \mathbb{R}^m} \left[\phi^*(-w) + \sup_{\beta \in \Omega} \langle w,\, F(\beta)\rangle\right]$$

Here, $\langle w,\, F(\beta)\rangle = \sum_{i=1}^m w_i P(y_i|\beta)$.

$$(\mathrm{CQ}) \qquad \exists\ \beta \in \Omega \text{ s.t. } P(y_i|\beta) > 0,\ i = 1, \ldots, m.$$

(CQ) implies strong duality,

Duality Theorem for $(\mathtt{P}) - (\mathtt{D})$

$$\boxed{(\mathtt{P}) \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}}(\mu)} \quad \boxed{(\mathtt{D}) \qquad \min_{w \in \mathbb{R}^m} [\psi^*(-w) + \delta^*_{\mathcal{P}}(S^* w)]}$$

Duality Theorem for $(\mathtt{P}) - (\mathtt{D})$

$$\boxed{(\mathtt{P}) \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}}(\mu)} \qquad \boxed{(\mathtt{D}) \qquad \min_{w \in \mathbb{R}^m} [\psi^*(-w) + \delta^*_{\mathcal{P}}(S^* w)]}$$

Constraint Qualification (CQ) for $(\mathtt{P}) - (\mathtt{D})$

$$\boxed{\mathrm{ri}\,(S\,[\mathcal{P}(\Omega)]) \cap \mathrm{ri}\,(\mathrm{dom}\,(\psi)) \neq \emptyset}$$

Duality Theorem for $(\mathtt{P}) - (\mathtt{D})$

$$\boxed{(\mathtt{P}) \qquad \min_{\mu \in \mathcal{B}(\Omega)} \psi(S\mu) + \delta_{\mathcal{P}}(\mu)} \qquad \boxed{(\mathtt{D}) \qquad \min_{w \in \mathbb{R}^m} [\psi^*(-w) + \delta^*_{\mathcal{P}}(S^* w)]}$$

Constraint Qualification (CQ) for $(\mathtt{P}) - (\mathtt{D})$

$$\boxed{\mathrm{ri}\,(S\,[\mathcal{P}(\Omega)]) \cap \mathrm{ri}\,(\mathrm{dom}\,(\psi)) \neq \emptyset}$$

Strong Duality Theorem: If CQ holds, then there exists an optimal $(\mathtt{P})-(\mathtt{D})$ pair (μ, w) at which

$$\psi(S\mu) + \delta_{\mathcal{P}}(\mu)] + [\psi^*(-w) + \delta^*_{\mathcal{P}}(S^* w)] = 0.$$

(P) $\implies$ Convex-Composite Optimization

Ingredients from Choquet Theory.

- $x \in \mathrm{ext}(C)$ (extreme points of C) if

$$[x = (1-\lambda)z + \lambda y,\ \ z, y \in C \text{ with } \lambda \in (0,1)] \implies [x = z = y].$$

$(\mathrm{P}) \implies$ Convex-Composite Optimization

Ingredients from Choquet Theory.

- $x \in \mathrm{ext}(C)$ (extreme points of C) if

$$[x = (1-\lambda)z + \lambda y,\ \ z, y \in C \text{ with } \lambda \in (0,1)] \implies [x = z = y].$$

- Krein-Milman Theorem: *A compact convex subset C of a Hausdorff locally convex topological vector space is equal to the closed convex hull of its extreme points, $C = \overline{\mathrm{co}}(\mathrm{ext}(C))$.*

$(\mathrm{P}) \implies$ Convex-Composite Optimization

Ingredients from Choquet Theory.

- $x \in \mathrm{ext}(C)$ (extreme points of C) if

$$[x = (1-\lambda)z + \lambda y,\ z, y \in C \text{ with } \lambda \in (0,1)] \implies [x = z = y].$$

- Krein-Milman Theorem: *A compact convex subset C of a Hausdorff locally convex topological vector space is equal to the closed convex hull of its extreme points, $C = \overline{\mathrm{co}}(\mathrm{ext}(C))$.*

The closure is not required in finite dimensions.

Ingredients from Choquet Theory.

- $x \in \mathrm{ext}(C)$ (extreme points of C) if

$$[x = (1-\lambda)z + \lambda y,\ \ z, y \in C \text{ with } \lambda \in (0,1)] \implies [x = z = y].$$

- Krein-Milman Theorem: *A compact convex subset C of a Hausdorff locally convex topological vector space is equal to the closed convex hull of its extreme points, $C = \overline{\mathrm{co}}(\mathrm{ext}(C))$.*
The closure is not required in finite dimensions.

- Let X and Y be two Hausdorff locally convex topological vector spaces, $C \subset X$ compact convex, and let $T \in \mathcal{L}[X, Y]$. Then TC is compact convex and $\mathrm{ext}(TC) \subset T\mathrm{ext}(C)$.

(P) $\implies$ Convex-Composite Optimization

Ingredients from Choquet Theory.

- $x \in \mathrm{ext}(C)$ (extreme points of C) if

$$[x = (1-\lambda)z + \lambda y,\ \ z, y \in C \text{ with } \lambda \in (0,1)] \implies [x = z = y].$$

- Krein-Milman Theorem: *A compact convex subset C of a Hausdorff locally convex topological vector space is equal to the closed convex hull of its extreme points, $C = \overline{\mathrm{co}}(\mathrm{ext}(C))$.*
The closure is not required in finite dimensions.

- Let X and Y be two Hausdorff locally convex topological vector spaces, $C \subset X$ compact convex, and let $T \in \mathcal{L}[X, Y]$. Then TC is compact convex and $\mathrm{ext}(TC) \subset T\mathrm{ext}(C)$.

- $\mathcal{B}(\Omega)$ is a Hausdorff locally convex topological vector space and $\mathcal{P}(\Omega)$ is a w*-compact convex subset, so

$$S\mathcal{P}(\Omega) = \mathrm{co}(S[\mathrm{ext}(\mathcal{P}(\Omega))]).$$

The Extreme points of $\mathcal{P}(\Omega)$

- $\mathrm{ext}(\mathcal{P}(\Omega)) = \{\mathsf{a}_\beta \mid \beta \in \Omega\}$, the set of Dirac measures on Ω, where for all $A \subset \Omega$ (Borel),

$$\mathsf{a}_\beta(A) := \begin{cases} 1 & , \ \beta \in A, \\ 0 & , \ \beta \notin A. \end{cases}$$

-

The Extreme points of $\mathcal{P}(\Omega)$

- $\mathrm{ext}(\mathcal{P}(\Omega)) = \{\mathsf{a}_\beta \mid \beta \in \Omega\}$, the set of Dirac measures on Ω, where for all $A \subset \Omega$ (Borel),

$$\mathsf{a}_\beta(A) := \begin{cases} 1 & , \ \beta \in A, \\ 0 & , \ \beta \notin A. \end{cases}$$

- The representation of the linear transformation S yields

$$S\mathsf{a}_{\bar{\beta}} = \int_\Omega F(\beta) d\mathsf{a}_{\bar{\beta}}(\beta) = F(\bar{\beta}).$$

The Extreme points of $\mathcal{P}(\Omega)$

- $\operatorname{ext}(\mathcal{P}(\Omega)) = \{\mathsf{a}_\beta \mid \beta \in \Omega\}$, the set of Dirac measures on Ω, where for all $A \subset \Omega$ (Borel),

$$\mathsf{a}_\beta(A) := \begin{cases} 1 & , \ \beta \in A, \\ 0 & , \ \beta \notin A. \end{cases}$$

- The representation of the linear transformation S yields

$$S\mathsf{a}_{\bar{\beta}} = \int_\Omega F(\beta) d\mathsf{a}_{\bar{\beta}}(\beta) = F(\bar{\beta}).$$

- Consequently,

$$S[\operatorname{ext}(\mathcal{P}(\Omega))] = \{S\mathsf{a}_\beta \mid \beta \in \Omega\} = \{F(\beta) \mid \beta \in \Omega\}.$$

Reduction of $(\mathtt{P})$ to Finite Dimensions

$$\min_{w} \{ \psi(w) \mid w \in S\mathcal{P}(\Omega) \} = \min_{w} \{ \psi(w) \mid w \in \mathrm{co}(S\,[\mathrm{ext}(\mathcal{P}(\Omega))]) \}$$

Reduction of $(\mathtt{P})$ to Finite Dimensions

$$
\begin{aligned}
\min_w \{\psi(w) \mid w \in S\mathcal{P}(\Omega)\} &= \min_w \{\psi(w) \mid w \in \operatorname{co}(S[\operatorname{ext}(\mathcal{P}(\Omega))])\} \\
&= \min_w \{\psi(w) \mid w \in \operatorname{co}(\{F(\beta) \mid \beta \in \Omega\})\}
\end{aligned}
$$

Reduction of (P) to Finite Dimensions

$$\begin{aligned}\min_w \{\psi(w) \mid w \in S\mathcal{P}(\Omega)\} &= \min_w \{\psi(w) \mid w \in \mathrm{co}(S\,[\mathrm{ext}(\mathcal{P}(\Omega))])\} \\ &= \min_w \{\psi(w) \mid w \in \mathrm{co}(\{F(\beta) \mid \beta \in \Omega\})\}\end{aligned}$$

(by Carathéodory's Theorem with $\hat{m} > m$)

Reduction of $(\mathtt{P})$ to Finite Dimensions

$$\min_w \{\psi(w) \mid w \in S\mathcal{P}(\Omega)\} = \min_w \{\psi(w) \mid w \in \mathrm{co}(S\,[\mathrm{ext}(\mathcal{P}(\Omega))])\}$$
$$= \min_w \{\psi(w) \mid w \in \mathrm{co}(\{F(\beta) \mid \beta \in \Omega\})\}$$

(by Carathéodory's Theorem with $\hat{m} > m$)

$$= \min_{\lambda \in \Delta_{\hat{m}},\ \beta^i} \left\{\psi\left(\mathcal{F}(\beta^1, \dots, \beta^{\hat{m}})\lambda\right) \,\middle|\, \beta^i \in \Omega,\ i = 1, \dots, \hat{m}\right\}$$

where $\Delta_{\hat{m}} := \left\{\lambda \in \mathbb{R}^{\hat{m}}_+ \,\middle|\, \mathsf{e}^T\lambda = 1\right\}$ (the unit simplex), and e denotes the vector of all ones, and

$$\mathcal{F}(\beta^1, \dots, \beta^m) = [F(\beta^1), \dots, F(\beta^{\hat{m}})],$$

Reduction of $(\mathtt{P})$ to Finite Dimensions

$$\min_w \{\psi(w) \mid w \in S\mathcal{P}(\Omega)\} = \min_w \{\psi(w) \mid w \in \mathrm{co}(S\,[\mathrm{ext}(\mathcal{P}(\Omega))])\}$$
$$= \min_w \{\psi(w) \mid w \in \mathrm{co}(\{F(\beta) \mid \beta \in \Omega\})\}$$

(by Carathéodory's Theorem with $\hat{m} > m$)

$$= \min_{\lambda \in \Delta_{\hat{m}},\ \beta^i} \left\{ \psi\left(\mathcal{F}(\beta^1, \ldots, \beta^{\hat{m}})\lambda\right) \,\middle|\, \beta^i \in \Omega,\ i = 1, \ldots, \hat{m} \right\} \sim \widehat{(\mathtt{P})}$$

where $\Delta_{\hat{m}} := \left\{ \lambda \in \mathbb{R}^{\hat{m}}_+ \,\middle|\, \mathsf{e}^T\lambda = 1 \right\}$ (the unit simplex), and e denotes the vector of all ones, and

$$\mathcal{F}(\beta^1, \ldots, \beta^m) = [F(\beta^1), \ldots, F(\beta^{\hat{m}})],$$

Reduction of (P) to Finite Dimensions

$$\begin{aligned}\min_w \{\psi(w) \mid w \in S\mathcal{P}(\Omega)\} &= \min_w \{\psi(w) \mid w \in \mathrm{co}(S[\mathrm{ext}(\mathcal{P}(\Omega))])\} \\ &= \min_w \{\psi(w) \mid w \in \mathrm{co}(\{F(\beta) \mid \beta \in \Omega\})\}\end{aligned}$$

(by Carathéodory's Theorem with $\hat{m} > m$)

$$= \min_{\lambda \in \Delta_{\hat{m}},\ \beta^i} \left\{ \psi\left(\mathcal{F}(\beta^1, \ldots, \beta^{\hat{m}})\lambda\right) \,\middle|\, \beta^i \in \Omega,\ i = 1, \ldots, \hat{m} \right\} \sim \widehat{(\mathrm{P})}$$

where $\Delta_{\hat{m}} := \left\{\lambda \in \mathbb{R}^{\hat{m}}_+ \,\middle|\, \mathsf{e}^T\lambda = 1\right\}$ (the unit simplex), and e denotes the vector of all ones, and

$$\mathcal{F}(\beta^1, \ldots, \beta^m) = [F(\beta^1), \ldots, F(\beta^{\hat{m}})],$$

$\widehat{(\mathrm{P})}$ is convex-composite but not Convex!

Reduction of $(\mathtt{P})$ to Finite Dimensions

$$\min_w \{\psi(w) \mid w \in S\mathcal{P}(\Omega)\} = \min_w \{\psi(w) \mid w \in \mathrm{co}(S\,[\mathrm{ext}(\mathcal{P}(\Omega))])\}$$
$$= \min_w \{\psi(w) \mid w \in \mathrm{co}(\{F(\beta) \mid \beta \in \Omega\})\}$$

(by Carathéodory's Theorem with $\hat{m} > m$)

$$= \min_{\lambda \in \Delta_{\hat{m}},\ \beta^i} \left\{\psi\left(\mathcal{F}(\beta^1, \ldots, \beta^{\hat{m}})\lambda\right) \,\Big|\, \beta^i \in \Omega,\ i = 1, \ldots, \hat{m}\right\} \sim \widehat{(\mathtt{P})}$$

where $\Delta_{\hat{m}} := \left\{\lambda \in \mathbb{R}^{\hat{m}}_+ \,\middle|\, \mathsf{e}^T\lambda = 1\right\}$ (the unit simplex), and e denotes the vector of all ones, and

$$\mathcal{F}(\beta^1, \ldots, \beta^m) = [F(\beta^1), \ldots, F(\beta^{\hat{m}})],$$

$\widehat{(\mathtt{P})}$ is convex-composite but not Convex!

But the dual of $\widehat{(\mathtt{P})}$ is $(\mathtt{D})$!

Numerical Methods

1) EM methods

2) mesh or grid (including random mesh generation) and moving mesh methods

3) (Steepest descent) Frank Wolfe methods, vertex direction methods, cutting plane methods
(Mallet 86', Böhning 85'-86')

4) smoothing and convex composite methods

5) projected subgradient descent methods

6) Bender's decomposition methods

Vertex direction methods for NPML

Given observation $y^1, \ldots, y^m \in \mathbb{R}^N$ solve

$$\min_{\mu \in \mathcal{P}(\Omega)} L(\mu) := \varphi\left(\int_\Omega F(\beta)\mu(d\beta)\right)$$

where $K \subset \mathbb{E}$ is closed convex and

$$F(\beta) = \begin{pmatrix} P(y^1|\beta) \\ \vdots \\ P(y^m|\beta) \end{pmatrix} \text{ and } \varphi(z) = \begin{cases} -\sum_{i=1}^m \log(z_i) & , \ z \in \mathbb{R}^m_{++}, \\ +\infty & , \text{ else.} \end{cases}$$

Vertex direction methods for NPML

Given observation $y^1, \ldots, y^m \in \mathbb{R}^N$ solve

$$\min_{\mu \in \mathcal{P}(\Omega)} L(\mu) := \varphi\left(\int_\Omega F(\beta)\mu(d\beta)\right)$$

where $K \subset \mathbb{E}$ is closed convex and

$$F(\beta) = \begin{pmatrix} P(y^1|\beta) \\ \vdots \\ P(y^m|\beta) \end{pmatrix} \text{ and } \varphi(z) = \begin{cases} -\sum_{i=1}^m \log(z_i) & , \ z \in \mathbb{R}^m_{++}, \\ +\infty & , \ \text{else.} \end{cases}$$

Finite dimensional version:

$$\min_{w \in C} \varphi(w),$$

where $C = S\mathcal{P}(\Omega) = \mathrm{co}(\{F(\beta) \mid \beta \in \Omega\})$.

First-order optimality conditions

$$(\mathtt{P}) \quad \min_{w \in C} \varphi(w), \quad C = \mathrm{co}(\{F(\beta) \mid \beta \in \Omega\})$$

We have $\nabla\varphi(\bar{w}) = -\bar{w}^{-1}$ where $(\bar{w}^{-1})_i = 1/\bar{w}_i$ is the componentwise inverse if the vector $\bar{w}$.

$$\bar{w} \in \mathbb{R}^m_{++} \text{ solves } (\mathtt{P})$$

$$\Longleftrightarrow$$

$$\varphi'(\bar{w}; w - \bar{w}) \geq 0 \quad \forall\, w \in C$$

First-order optimality conditions

$$(\mathtt{P}) \quad \min_{w \in C} \varphi(w), \quad C = \mathrm{co}(\{F(\beta) \mid \beta \in \Omega\})$$

We have $\nabla\varphi(\bar{w}) = -\bar{w}^{-1}$ where $(\bar{w}^{-1})_i = 1/\bar{w}_i$ is the componentwise inverse if the vector $\bar{w}$.

$$\bar{w} \in \mathbb{R}^m_{++} \text{ solves } (\mathtt{P})$$

$$\Longleftrightarrow$$

$$\varphi'(\bar{w}; w - \bar{w}) \geq 0 \quad \forall\, w \in C$$

$$\Longleftrightarrow$$

$$m \geq \left\langle \bar{w}^{-1},\, w \right\rangle \geq 0 \quad \forall\, w \in C$$

$$\Longleftrightarrow$$

$$m \geq \sup_{\beta \in \Omega} \left\langle \bar{w}^{-1},\, F(\beta) \right\rangle$$

First-order optimality conditions

$$\texttt{(P)} \quad \min_{w \in C} \varphi(w), \quad C = \operatorname{co}(\{F(\beta) \mid \beta \in \Omega\})$$

We have $\nabla\varphi(\bar{w}) = -\bar{w}^{-1}$ where $(\bar{w}^{-1})_i = 1/\bar{w}_i$ is the componentwise inverse if the vector $\bar{w}$.

$$\bar{w} \in \mathbb{R}^m_{++} \text{ solves } \texttt{(P)}$$

$$\Longleftrightarrow$$

$$\varphi'(\bar{w}; w - \bar{w}) \geq 0 \quad \forall\, w \in C$$

$$\Longleftrightarrow$$

$$m \geq \left\langle \bar{w}^{-1},\, w \right\rangle \geq 0 \quad \forall\, w \in C$$

$$\Longleftrightarrow$$

$$m \geq \sup_{\beta \in \Omega} \left\langle \bar{w}^{-1},\, F(\beta) \right\rangle$$

$$\Longleftrightarrow$$

$$m = \sup_{\beta \in \Omega} \left\langle \bar{w}^{-1},\, F(\beta) \right\rangle$$

Steepest descent: Vertex direction method

Given

$$\bar{w} := \sum_{j=1}^{\hat{m}} \lambda_j F(\beta^j) = S\left[\sum_{j=1}^{\hat{m}} \lambda_j \mathsf{a}_{\beta^j}\right], \quad \lambda \in \Delta_{\hat{m}},$$

solve

$$\inf_{w \in S\mathcal{P}(\Omega)} \phi'(\bar{w}; w - \bar{w})$$

Steepest descent: Vertex direction method

Given

$$\bar{w} := \sum_{j=1}^{\hat{m}} \lambda_j F(\beta^j) = S\left[\sum_{j=1}^{\hat{m}} \lambda_j \mathsf{a}_{\beta^j}\right], \quad \lambda \in \Delta_{\hat{m}},$$

solve

$$\inf_{w \in S\mathcal{P}(\Omega)} \phi'(\bar{w}; w - \bar{w}) = \inf_{w \in S\mathcal{P}(\Omega)} \left\langle -\bar{w}^{-1},\, w - \bar{w} \right\rangle$$

Steepest descent: Vertex direction method

Given

$$\bar{w} := \sum_{j=1}^{\hat{m}} \lambda_j F(\beta^j) = S\left[\sum_{j=1}^{\hat{m}} \lambda_j \mathsf{a}_{\beta^j}\right], \quad \lambda \in \Delta_{\hat{m}},$$

solve

$$\begin{aligned}
\inf_{w \in S\mathcal{P}(\Omega)} \phi'(\bar{w}; w - \bar{w}) &= \inf_{w \in S\mathcal{P}(\Omega)} \left\langle -\bar{w}^{-1}, \, w - \bar{w} \right\rangle \\
&= m - \sup_{\mu \in \mathcal{P}(\Omega)} \left\langle \bar{w}^{-1}, \, \int_\Omega F(\beta) d\mu \right\rangle \\
&= m - \sup_{\beta \in \Omega} \left\langle \bar{w}^{-1}, \, F(\beta) \right\rangle
\end{aligned}$$

Steepest descent: Vertex direction method

Given

$$\bar{w} := \sum_{j=1}^{\hat{m}} \lambda_j F(\beta^j) = S\left[\sum_{j=1}^{\hat{m}} \lambda_j \mathsf{a}_{\beta^j}\right], \quad \lambda \in \Delta_{\hat{m}},$$

solve

$$\begin{aligned}\inf_{w \in S\mathcal{P}(\Omega)} \phi'(\bar{w}; w - \bar{w}) &= \inf_{w \in S\mathcal{P}(\Omega)} \left\langle -\bar{w}^{-1}, w - \bar{w} \right\rangle \\ &= m - \sup_{\mu \in \mathcal{P}(\Omega)} \left\langle \bar{w}^{-1}, \int_\Omega F(\beta) d\mu \right\rangle \\ &= m - \sup_{\beta \in \Omega} \left\langle \bar{w}^{-1}, F(\beta) \right\rangle\end{aligned}$$

for β^+ and set $w^+ = F(\beta^+)$.

Steepest descent: Vertex direction method

Given

$$\bar{w} := \sum_{j=1}^{\hat{m}} \lambda_j F(\beta^j) = S\left[\sum_{j=1}^{\hat{m}} \lambda_j \mathsf{a}_{\beta^j}\right], \quad \lambda \in \Delta_{\hat{m}},$$

solve

$$\begin{aligned}
\inf_{w \in S\mathcal{P}(\Omega)} \phi'(\bar{w}; w - \bar{w}) &= \inf_{w \in S\mathcal{P}(\Omega)} \left\langle -\bar{w}^{-1}, w - \bar{w} \right\rangle \\
&= m - \sup_{\mu \in \mathcal{P}(\Omega)} \left\langle \bar{w}^{-1}, \int_\Omega F(\beta) d\mu \right\rangle \\
&= m - \sup_{\beta \in \Omega} \left\langle \bar{w}^{-1}, F(\beta) \right\rangle
\end{aligned}$$

for β^+ and set $w^+ = F(\beta^+)$.

Line search:

$$\min_{\tau > 0} \varphi(\bar{w} + \tau(w^+ - \bar{w}))$$

Steepest descent: Vertex direction method

Given

$$\bar{w} := \sum_{j=1}^{\hat{m}} \lambda_j F(\beta^j) = S\left[\sum_{j=1}^{\hat{m}} \lambda_j \mathsf{a}_{\beta^j}\right], \quad \lambda \in \Delta_{\hat{m}},$$

solve

$$\begin{aligned}
\inf_{w \in S\mathcal{P}(\Omega)} \phi'(\bar{w}; w - \bar{w}) &= \inf_{w \in S\mathcal{P}(\Omega)} \left\langle -\bar{w}^{-1},\, w - \bar{w} \right\rangle \\
&= m - \sup_{\mu \in \mathcal{P}(\Omega)} \left\langle \bar{w}^{-1},\, \int_\Omega F(\beta) d\mu \right\rangle \\
&= m - \sup_{\beta \in \Omega} \left\langle \bar{w}^{-1},\, F(\beta) \right\rangle
\end{aligned}$$

for β^+ and set $w^+ = F(\beta^+)$.
Line search:

$$\min_{\tau > 0} \varphi(\bar{w} + \tau(w^+ - \bar{w}))$$

Updated to the Vertex Exchange Methods. But this class of methods are quite slow.

Grid and sampling methods

Suppose $\hat{m} >> m$, that is $\{\beta^1, \ldots, \beta^{\hat{m}}\}$ is a grid on Ω, or a large sample of elements from Ω. Set $\Psi := \mathcal{F}(\beta^1, \ldots, \beta^{\hat{m}}) \in \mathbb{R}^{m \times \hat{m}}$ and consider the problem

$$\widehat{(\mathtt{P})}_{\hat{m}} \quad \min_{\lambda \in \Delta_{\hat{m}}} \varphi(\Psi\lambda).$$

Grid and sampling methods

Suppose $\hat{m} >> m$, that is $\{\beta^1, \ldots, \beta^{\hat{m}}\}$ is a grid on Ω, or a large sample of elements from Ω. Set $\Psi := \mathcal{F}(\beta^1, \ldots, \beta^{\hat{m}}) \in \mathbb{R}^{m\times\hat{m}}$ and consider the problem

$$\widehat{(\mathtt{P})}_{\hat{m}} \qquad \min_{\lambda\in\Delta_{\hat{m}}} \varphi(\Psi\lambda).$$

One can show that $\widehat{(\mathtt{P})}_{\hat{m}}$ is equivalent to

$$\min_{\lambda\geq 0} \varphi(\Psi\lambda) + \hat{m}(\mathsf{e}^T\lambda - 1).$$

Grid and sampling methods

Suppose $\hat{m} >> m$, that is $\{\beta^1, \ldots, \beta^{\hat{m}}\}$ is a grid on Ω, or a large sample of elements from Ω. Set $\Psi := \mathcal{F}(\beta^1, \ldots, \beta^{\hat{m}}) \in \mathbb{R}^{m \times \hat{m}}$ and consider the problem

$$\widehat{(\mathtt{P})}_{\hat{m}} \quad \min_{\lambda \in \Delta_{\hat{m}}} \varphi(\Psi\lambda).$$

One can show that $\widehat{(\mathtt{P})}_{\hat{m}}$ is equivalent to

$$\min_{\lambda \geq 0} \varphi(\Psi\lambda) + \hat{m}(\mathsf{e}^T\lambda - 1).$$

The log-barrier relaxation of this problem is

$$\widehat{(\mathtt{P})}_{\hat{m}}^{\tau} \quad \min_{\lambda} \varphi(\Psi\lambda) + \hat{m}(\mathsf{e}^T\lambda - 1) + \tau\varphi(\lambda).$$

Grid and sampling methods

$$\widehat{(\mathrm{P})}_{\hat{m}}^{\tau} \quad \min_{\lambda} \varphi(\Psi\lambda) + \hat{m}(\mathsf{e}^T\lambda - 1) + \tau\varphi(\lambda).$$

where $\bar{\lambda} \in \mathbb{R}^{\hat{m}}_{++}$ solves $\widehat{(\mathrm{P})}_{\hat{m}}^{\tau}$ it and only if there exist $z, w \in \mathbb{R}^{m}_{++},\ y \in \mathbb{R}^{\hat{m}}_{++}$ such that

$$\begin{aligned} \hat{m}\mathsf{e} &= \Psi^T w + y \\ z &= \Psi\bar{\lambda} \\ \mathsf{e} &= \mathsf{Diag}(w)\mathsf{Diag}(z)\mathsf{e} \\ \tau\mathsf{e} &= \mathsf{Diag}(\lambda)\mathsf{Diag}(y)\mathsf{e}\ . \end{aligned}$$

Grid and sampling methods

$$\widehat{(\mathrm{P})}_{\hat{m}}^{\tau} \quad \min_{\lambda} \varphi(\Psi\lambda) + \hat{m}(\mathsf{e}^T\lambda - 1) + \tau\varphi(\lambda).$$

where $\bar{\lambda} \in \mathbb{R}^{\hat{m}}_{++}$ solves $\widehat{(\mathrm{P})}_{\hat{m}}^{\tau}$ it and only if there exist $z, w \in \mathbb{R}^{m}_{++},\ y \in \mathbb{R}^{\hat{m}}_{++}$ such that

$$\begin{aligned} \hat{m}\mathsf{e} &= \Psi^T w + y \\ z &= \Psi\bar{\lambda} \\ \mathsf{e} &= \mathsf{Diag}(w)\mathsf{Diag}(z)\mathsf{e} \\ \tau\mathsf{e} &= \mathsf{Diag}(\lambda)\mathsf{Diag}(y)\mathsf{e}\ . \end{aligned}$$

Now apply an interior point predictor-corrector strategy ($\tau \downarrow 0$) to solve $\widehat{(\mathrm{P})}_{\hat{m}}$ quickly and accurately.

Grid and sampling methods

$$\widehat{(\mathrm{P})}^{\tau}_{\hat{m}} \quad \min_{\lambda} \varphi(\Psi\lambda) + \hat{m}(\mathsf{e}^T\lambda - 1) + \tau\varphi(\lambda).$$

where $\bar{\lambda} \in \mathbb{R}^{\hat{m}}_{++}$ solves $\widehat{(\mathrm{P})}^{\tau}_{\hat{m}}$ it and only if there exist $z, w \in \mathbb{R}^{m}_{++},\ y \in \mathbb{R}^{\hat{m}}_{++}$ such that

$$\begin{aligned}
\hat{m}\mathsf{e} &= \Psi^T w + y \\
z &= \Psi\bar{\lambda} \\
\mathsf{e} &= \mathsf{Diag}(w)\mathsf{Diag}(z)\mathsf{e} \\
\tau\mathsf{e} &= \mathsf{Diag}(\lambda)\mathsf{Diag}(y)\mathsf{e}\ .
\end{aligned}$$

Now apply an interior point predictor-corrector strategy ($\tau \downarrow 0$) to solve $\widehat{(\mathrm{P})}_{\hat{m}}$ quickly and accurately. $(3 \le m \le 15,\ 20,000 \le \hat{m} \le 80,000)$

An Algorithm for Nonparametric Estimation of A Multivariate Mixing Distribution with Applications to Population Pharmacokinetics, W.M.Yamada, M.N.Neely, J. Bartroff, D.S.Bayard, J.V. Burke, M. van Guilder, R.W.Jelliffe, A.Kryshchenko, R.Leary, T.Tatarinova, A.Schumitzky. Pharmaceutics. 2020 Dec 30;13(1):42.

EM Algorithm

The basic problem in λ:

$$\widehat{(\mathtt{P})}_{\mathsf{NPML}} \qquad \min_{\lambda \in \Delta_{\hat{m}}} \varphi(\Psi\lambda).$$

EM Algorithm

The basic problem in λ:

$$\widehat{(\mathrm{P})}_{\mathsf{NPML}} \qquad \min_{\lambda \in \Delta_{\hat{m}}} \varphi(\Psi\lambda).$$

Basic Assumptions for Ψ:

$$\Psi\Delta_{\hat{M}} \subset \mathbb{R}^M_+ \quad \text{and} \quad \exists\, \lambda \in \Delta_{\hat{M}} \quad \text{such that} \quad \Psi\lambda > 0\ .$$

EM Algorithm

The basic problem in λ:

$$\widehat{(\mathrm{P})}_{\text{NPML}} \quad \min_{\lambda \in \Delta_{\hat{m}}} \varphi(\Psi\lambda).$$

Basic Assumptions for Ψ:

$$\Psi\Delta_{\hat{M}} \subset \mathbb{R}_+^M \ \text{ and } \ \exists\, \lambda \in \Delta_{\hat{M}} \ \text{ such that } \ \Psi\lambda > 0 \ .$$

EM fixed point iteration:

$$\lambda^{\nu+1} = \frac{1}{M}\Lambda_\nu \Psi^T (\Psi\lambda^\nu)^{-1} \text{ with } \lambda > 0 \text{ and } \Psi\lambda_0 > 0.$$

EM Algorithm

The basic problem in λ:

$$\widehat{(\mathrm{P})}_{\mathsf{NPML}} \qquad \min_{\lambda \in \Delta_{\hat{m}}} \varphi(\Psi\lambda).$$

Basic Assumptions for Ψ:

$$\Psi\Delta_{\hat{M}} \subset \mathbb{R}^M_+ \ \text{ and } \ \exists\, \lambda \in \Delta_{\hat{M}} \ \text{ such that } \ \Psi\lambda > 0\ .$$

EM fixed point iteration:

$$\lambda^{\nu+1} = \frac{1}{M}\Lambda_\nu \Psi^T (\Psi\lambda^\nu)^{-1} \text{ with } \lambda > 0 \text{ and } \Psi\lambda_0 > 0.$$

Convergence: $\{\lambda^\nu\} \subset \Delta_{\hat{m}}$ and every cluster point solves $\widehat{(\mathrm{P})}_{\mathsf{NPML}}$.

Bender's decomposition

$$\min_{\lambda \in \Delta_{\hat{m}},\ x \in \Omega^{\hat{m}}} \varphi\left(\mathcal{F}(x)\lambda\right) \;=\; \min_{x \in \Omega^{\hat{m}}} \left[\min_{\lambda \in \Delta_{\hat{m}}} \varphi(\mathcal{F}(x)\lambda) \right]$$

Bender's decomposition

$$\min_{\lambda\in\Delta_{\hat{m}},\ x\in\Omega^{\hat{m}}} \varphi\left(\mathcal{F}(x)\lambda\right) \;=\; \min_{x\in\Omega^{\hat{m}}}\left[\min_{\lambda\in\Delta_{\hat{m}}} \varphi(\mathcal{F}(x)\lambda)\right]$$

Choose a smoothing function φ_τ for $\varphi + \delta_{\Delta_{\hat{m}}}$ and solve for decreasing τ:

$$\min_{x\in\Omega^{\hat{m}}} v_\tau(x), \quad \text{where } v_\tau(x) := \min_{\lambda} \varphi_\tau(\mathcal{F}(x)\lambda).$$

Bender's decomposition

$$\min_{\lambda\in\Delta_{\hat{m}},\ x\in\Omega^{\hat{m}}} \varphi\left(\mathcal{F}(x)\lambda\right) \;=\; \min_{x\in\Omega^{\hat{m}}}\left[\min_{\lambda\in\Delta_{\hat{m}}}\varphi(\mathcal{F}(x)\lambda)\right]$$

Choose a smoothing function φ_τ for $\varphi + \delta_{\Delta_{\hat{m}}}$ and solve for decreasing τ:

$$\min_{x\in\Omega^{\hat{m}}} v_\tau(x), \quad \text{where } v_\tau(x) := \min_{\lambda} \varphi_\tau(\mathcal{F}(x)\lambda).$$

$v_\tau \in \mathcal{C}^2$ and $v_\tau(x) = \min_{\lambda\in\Delta_{\hat{m}}} \varphi(\mathcal{F}(x)\lambda)$, $\nabla v_\tau(x)$, and $\nabla^2 v_\tau(x)$ can be rapidly and accurately evaluated.

Many unresolved key statistical questions

$$\int_{\Omega} P(y|\beta)\mu(d\beta)$$

a mixture density with mixing measure $\mu \in \mathcal{P}(\Omega)$.

Many unresolved key statistical questions

$$\int_{\Omega} P(y|\beta)\mu(d\beta)$$

a mixture density with mixing measure $\mu \in \mathcal{P}(\Omega)$.

Goal: Estimate μ from observations of $y^1, \ldots, y^m \in \mathbb{R}^N$.

Many unresolved key statistical questions

$$\int_{\Omega} P(y|\beta)\mu(d\beta)$$

a mixture density with mixing measure $\mu \in \mathcal{P}(\Omega)$.

Goal: Estimate μ from observations of $y^1, \ldots, y^m \in \mathbb{R}^N$.

Let μ_m be the maximum likelihood estimate.

- As $m \uparrow \infty$, does μ_m converge to something of interest?
- Does it converge to the measure describing the population distribution in the mixed effects in NPML model?
- If it does converge, in what sense does it converge, and are there error estimates?
- When do you have enough samples, or data?